MODIFYING SQP FOR DEGENERATE PROBLEMS

STEPHEN J. WRIGHT*

Abstract. Most local convergence analyses of the sequential quadratic programming (SQP) algorithm for nonlinear programming make strong assumptions about the solution, namely, that the active constraint gradients are linearly independent and that no weakly active constraints exist. In this paper, we establish a framework for variants of SQP that retain the characteristic superlinear convergence rate even when these assumptions are relaxed, proving general convergence results and placing some recently proposed SQP variants in this framework. We discuss the reasons for which implementations of SQP often continue to exhibit good local convergence behavior even when the assumptions commonly made in the analysis are violated. Finally, we describe a new algorithm that formalizes and extends standard SQP implementation techniques, and we prove convergence results for this method also.

AMS subject classifications. 90C33, 90C30, 49M45

1. Introduction. We investigate local convergence properties of variants of the sequential quadratic programming (SQP) algorithm applied to the nonlinear programming problem

$$(1.1) \qquad \text{NLP:} \qquad \min_z \phi(z) \qquad \text{subject to } g(z) \le 0,$$

where $\phi : \mathsf{R}^n \to \mathsf{R}$ and $g : \mathsf{R}^n \to \mathsf{R}^m$ are twice Lipschitz continuously differentiable functions. We are interested in *degenerate* problems: those for which the active constraint gradients at the solution are linearly dependent and/or the strict complementarity condition fails to hold.

We showed in [18] that even when strict complementarity, second-order sufficient conditions, and a constraint qualification hold, nonuniqueness of the optimal multiplier can produce nonsuperlinear behavior of SQP. Motivated by this observation and by the fact that primal-dual interior-point algorithms for related problems converge superlinearly under the conditions just described [19, 16], we proposed a stabilized SQP (sSQP) method [18] and proved a local superlinear convergence result, later enhanced by Hager [11]. Independently, Fischer [8] proposed an algorithm in which a special procedure for choosing the Lagrange multiplier estimate is inserted between iterations of SQP. He proved superlinear convergence under slightly different assumptions from ours.

Our purposes in this paper are twofold. First, we introduce a common framework, which we call iSQP (for inexact SQP) that allows for a unified analysis of the stabilization procedures of the preceding paragraph. We prove general convergence results for methods in the iSQP framework, highlighting the effect on the convergence rate of changes between successive Lagrange multiplier estimates.

Our second goal requires a little more explanation. Implementations of SQP (for example, SNOPT [10]) often continue to exhibit good local convergence behavior even on degenerate problems, in spite of the facts that such problems fail to satisfy the standard assumptions made in local convergence analyses and that theoretical examples of poor convergence behavior are easy to construct (see the example in Wright [18]). The iSQP framework proves to be useful in providing some theoretical

support for this good practical performance. We find that the strategy of using the active (or working) set from the QP subproblem at the previous iteration as the initial active set for the current iteration is important in explaining the good behavior, as is the fact that the solver of the QP subproblem is allowed to return a slightly infeasible answer. Further, we propose and analyze an algorithm called SQPsws, which in which the techniques used in existing implementations are formalized to produce an algorithm whose local convergence is superlinear under certain assumptions.

The main point of difference between the basic SQP algorithm as presented here and the versions that are implemented in standard software is that the implementations usually make use of quasi-Newton Hessian approximations, whereas we assume here that exact Hessians are used. Still, we believe that our observations below are relevant to the quasi-Newton case and, in particular, that quasi-Newton versions of the various algorithms discussed here would exhibit fast local convergence. Extension of the analysis to this case would, however, not be trivial, since it would have to take into account such factors as the effects of degeneracy on the quasi-Newton updates, so we leave this issue for possible future work. Another point of difference between our SQP algorithm and the implementations is that the implementations contain various algorithmic devices to ensure global convergence, which we ignore these here because of our focus on local properties.

The remainder of the paper is structured as follows. In Section 2, we outline first-order optimality conditions and define various terms and assumptions that are used in the remainder of the paper. Section 3 defines the various second-order sufficient conditions that are required by the algorithms described in later sections. The iSQP framework is defined in Section 4, where we also prove a useful result about the active set identified by the iSQP subproblem. Section 5 contains the main results about convergence of algorithms in the iSQP framework. Brief discussions of the stabilized SQP algorithm and Fischer's approach are given in Sections 6 and 7, respectively, where we outline how both methods fit into the iSQP framework. Finally, the new algorithm SQPsws is described and some superlinear convergence results are proved for it in Section 8.

2. Assumptions, Notation, and Basic Results. We now review the optimality conditions for (1.1) and discuss various assumptions that are used in subsequent sections. These include second-order sufficient conditions of various types, along with complementarity conditions and the Mangasarian-Fromovitz constraint qualification (MFCQ). Finally, we quote a result that plays a key role in the analysis of the remainder of the paper—that MFCQ is equivalent to boundedness of the set of optimal Langrange multipliers.

The Lagrangian for (1.1) is

$$(2.1) \qquad \mathcal{L}(z, \lambda) = \phi(z) + \lambda^T g(z),$$

where $\lambda \in \mathsf{R}^m$ is the vector of Lagrange multipliers. We assume throughout that z^* is a strict local solution of (1.1). When a constraint qualification holds at z^* (see discussion below), first-order necessary conditions imply that there exists a vector $\lambda^* \in \mathsf{R}^m$ such that

$$(2.2) \qquad \mathcal{L}_z(z^*, \lambda^*) = 0, \quad g(z^*) \leq 0, \quad \lambda^* \geq 0, \quad (\lambda^*)^T g(z^*) = 0.$$

These relations are the well-known Karush-Kuhn-Tucker (KKT) conditions. The following sets play an important role in the remainder of the paper:

$$(2.3a) \qquad \mathcal{S}_\lambda = \{\lambda^* \mid \lambda^* \text{ satisfies } (2.2)\}$$

(2.3b)
$$\mathcal{S} = \{z^*\} \times \mathcal{S}_\lambda.$$

We can write the conditions (2.2) alternatively as

(2.4)
$$\left[\begin{array}{c} \nabla \phi(z^*) + \nabla g(z^*)\lambda^* \\ g(z^*) \end{array} \right] \in \left[\begin{array}{c} 0 \\ N(\lambda^*) \end{array} \right],$$

where $N(\lambda)$ is the set defined by

(2.5)
$$N(\lambda) \stackrel{\text{def}}{=} \left\{ \begin{array}{ll} \{y \,|\, y \le 0 \text{ and } y^T\lambda = 0\} & \text{if } \lambda \ge 0, \\ \emptyset & \text{otherwise.} \end{array} \right.$$

The *active set* at z^* is defined by

(2.6)
$$\mathcal{B} = \{i = 1, 2, \ldots, m \,|\, g_i(z^*) = 0\}.$$

For any optimal multiplier $\lambda^* \in \mathcal{S}_\lambda$, we define the set $\mathcal{B}_+(\lambda^*)$ to be the "support" of λ^*, that is,

$$\mathcal{B}_+(\lambda^*) = \{i \in \mathcal{B} \,|\, \lambda_i^* > 0\}.$$

We define $\mathcal{B}_+$ (without argument) as

(2.7)
$$\mathcal{B}_+ \stackrel{\text{def}}{=} \cup_{\lambda^* \in \mathcal{S}_\lambda} \mathcal{B}_+(\lambda^*)$$

and denote its complement in $\mathcal{B}$ by $\mathcal{B}_0$, that is,

$$\mathcal{B}_0 \stackrel{\text{def}}{=} \mathcal{B} \backslash \mathcal{B}_+ .$$

Note that $\mathcal{B}_0$ is the set of indices $i \in \mathcal{B}$ such that $\lambda_i^* = 0$ for all $\lambda^* \in \mathcal{S}_\lambda$. The *strict complementarity* (SC) condition for the set $\mathcal{S}$ (which we use only sparingly in this paper) is that

(2.8)
$$\mathcal{B}_0 = \emptyset.$$

At some points in the paper, we use a condition that Fischer [8] calls *weak complementarity* (WCC), namely that

(2.9) $\text{Range}[\nabla g_i(z^*)]_{i \in \mathcal{B}_+(\lambda^*)} = \text{Range}[\nabla g_i(z^*)]_{i \in \mathcal{B}_+}, \qquad$ for all $\lambda^* \in \mathcal{S}_\lambda.$

Despite its name, WCC is not weaker than SC; neither condition implies the other.

In Section 4, we define the term *strict working set* to be, roughly speaking, the set of indices in $i \in \{1, 2, \ldots, m\}$ for which the Lagrange multipliers λ_i of the (possibly inexact) QP subproblem are strictly positive.

We assume throughout that the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at z^* [15]. That is,

(2.10)
$$\nabla g_{\mathcal{B}}(z^*)^T \bar{y} < 0 \quad \text{for some } \bar{y} \in \mathsf{R}^n,$$

where $\nabla g_{\mathcal{B}}(\cdot)$ is the $n \times |\mathcal{B}|$ matrix whose rows $\nabla g_i(\cdot)$, $i \in \mathcal{B}$, are the gradients of the functions g_i, $i \in \mathcal{B}$.

The general smoothness and first-order assumption that we make throughout the paper is as follows.

ASSUMPTION 1. *The functions $\phi(\cdot)$ and $g(\cdot)$ are twice Lipschitz continuously differentiable in an open neighborhood of z^*, and the first-order condition (2.2) is satisfied at z^*.*

The following result concerning boundedness of the optimal multiplier set $\mathcal{S}_\lambda$ is used often in the analysis of later sections.

LEMMA 2.1. (Gauvin [9]) *Suppose that Assumption 1 holds. Then $\mathcal{S}_\lambda$ defined in (2.3a) is bounded if and only if the MFCQ (2.10) is satisfied.*

Since $\mathcal{S}_\lambda$ is defined by the linear conditions $\nabla\phi(z^*) + \nabla g(z^*)\lambda^*$ and $\lambda^* \geq 0$, it is closed and convex. Therefore, under the conditions of Lemma 2.1, it is also compact.

We use the notation $\delta(\cdot)$ to denote Euclidean distances from the primal, dual, and primal-dual optimal sets, according to context. Specifically, we define

$$(2.11) \quad \delta(z) \stackrel{\text{def}}{=} \|z - z^*\|, \quad \delta(\lambda) \stackrel{\text{def}}{=} \text{dist}\,(\lambda, \mathcal{S}_\lambda), \quad \delta(z, \lambda) \stackrel{\text{def}}{=} \text{dist}\,((z, \lambda), \mathcal{S}).$$

We also use $P(\lambda)$ to denote the projection of λ onto $\mathcal{S}_\lambda$; that is, we have $P(\lambda) \in \mathcal{S}_\lambda$ and $\|P(\lambda) - \lambda\| = \text{dist}\,(\lambda, \mathcal{S}_\lambda)$. Note that from (2.11) we have $\delta(z)^2 + \delta(\lambda)^2 = \delta(z, \lambda)^2$, and therefore

$$(2.12) \qquad\qquad \delta(z) \leq \delta(z, \lambda), \qquad \delta(\lambda) \leq \delta(z, \lambda).$$

For further analysis of these errors, we use $\mathcal{B}$ and $\mathcal{B}_+$ to define a direction set $\mathcal{T}$ as follows:

$$(2.13) \qquad\qquad \mathcal{T} = \left\{ w \mid \begin{array}{l} \nabla g_i(z^*)^T w = 0 \text{ for } i \in \mathcal{B}_+ \\ \nabla g_i(z^*)^T w \leq 0 \text{ for } i \in \mathcal{B}_0 \end{array} \right\}.$$

$$e(z) \stackrel{\text{def}}{=} z - z^*$$

and decompose it as

$$(2.14) \qquad\qquad e(z) = e_{\mathcal{T}}(z) + e_{\mathcal{N}}(z),$$

where $e_{\mathcal{T}}(z)$ is the projection of $e(z)$ onto the cone $\mathcal{T}$ and $e_{\mathcal{N}}(z)$ is the remainder (which is, of course, normal to $\mathcal{T}$ at $e_{\mathcal{T}}(z)$). In fact, there are coefficients ξ_i, $i \in \mathcal{B}$ (not necessarily unique), such that

$$(2.15) \qquad e_{\mathcal{N}}(z) = \sum_{i \in \mathcal{B}_+} \xi_i \nabla g_i(z^*) + \sum_{i \in \mathcal{B}_0} \xi_i \nabla g_i(z^*), \qquad \xi_i \geq 0 \text{ for } i \in \mathcal{B}_0.$$

Since $\mathcal{T}$ is a cone, it is easy to see that $e_{\mathcal{T}}(\cdot)$ and $e_{\mathcal{N}}(\cdot)$ are continuous in their arguments and that

$$e_{\mathcal{N}}(\alpha z) = \alpha e_{\mathcal{N}}(z), \qquad e_{\mathcal{T}}(\alpha z) = \alpha e_{\mathcal{T}}(z), \qquad \text{for all } \alpha \geq 0.$$

Moreover, since $e_{\mathcal{N}}(z)^T e_{\mathcal{T}}(z) = 0$, we have

$$\|e_{\mathcal{N}}(z)\|^2 + \|e_{\mathcal{T}}(z)\|^2 = \|e(z)\|^2 = \delta(z)^2 \leq \delta(z, \lambda)^2$$

and therefore

$$(2.16) \qquad\qquad \|e_{\mathcal{N}}(z)\| \leq \delta(z), \qquad \|e_{\mathcal{T}}(z)\| \leq \delta(z).$$

We use order notation in the following (standard) way: If a matrix, vector, or scalar quantity M is a function of another matrix, vector, or scalar quantity A, we write $M = O(\|A\|)$ if there is a constant β such that $\|M\| \leq \beta\|A\|$ for all $\|A\|$ sufficiently small. We write $M = \Omega(\|A\|)$ if there is a constant β such that $\|M\| \geq \beta^{-1}\|A\|$ for all $\|A\|$ sufficiently small, and $M = \Theta(\|A\|)$ if both $M = O(\|A\|)$ and $M = \Omega(\|A\|)$. We write $M = o(\|A\|)$ if for all sequences $\{A_k\}$ with $\|A_k\| \to 0$, the corresponding sequence $\{M_k\}$ satisfies $\|M_k\|/\|A_k\| \to 0$.

If r is a vector and $\mathcal{A}$ is an index set, we use $r_\mathcal{A}$ to denote the subvector consisting of components r_i, $i \in \mathcal{A}$.

3. Second-Order Conditions. The presence of degeneracy allows for a variety of second-order sufficient conditions, all of which can be expected to hold for a wide range of problems and all of which are useful in investigating the local convergence properties of various algorithms. In this section, we define three such conditions that are needed by algorithms in later sections. We also introduce "extended" variants of the nonlinear programming problem (1.1) that differ from (1.1) only in that just a subset of the constraints is enforced. For some of these subsets, z^* remains a strict local solution satisfying some second-order sufficient condition; such subsets are particularly useful in the context of the algorithm to be discussed in Section 8. Finally, we include here several results that relate the conditions introduced in this section to the assumptions of the preceding section.

Second-order sufficient conditions typically assume that there is a positive value $\sigma > 0$ such that the condition

$$(3.1) \qquad w^T \mathcal{L}_{zz}(z^*, \lambda^*) w \geq \sigma \|w\|^2$$

holds, for some set of λ^* and w vectors. The three conditions used in this paper are as follows.

Condition 2s.1. (Second-order sufficient condition.) The condition (3.1) holds for all $\lambda^* \in \mathcal{S}_\lambda$ and all w such that

$$\nabla g_i(z^*)^T w = 0, \quad \text{for all } i \in \mathcal{B}_+,$$
$$\nabla g_i(z^*)^T w \leq 0, \quad \text{for all } i \in \mathcal{B}_0;$$

that is, $w \in \mathcal{T}$.

Condition 2s.2. (Strong second-order sufficient condition.) The condition (3.1) holds for all $\lambda^* \in \mathcal{S}_\lambda$ and all w such that

$$\nabla g_i(z^*)^T w = 0, \quad \text{for all } i \in \mathcal{B}_+.$$

Condition 2s.3. (Locally strong second-order sufficient condition.) For each $\lambda^* \in \mathcal{S}_\lambda$, the condition (3.1) holds for all w such that

$$(3.2) \qquad \nabla g_i(z^*)^T w = 0, \quad \text{for all } i \in \mathcal{B}_+(\lambda^*).$$

Any of these conditions, in tandem with Assumption 1, is sufficient to guarantee that z^* is a strict local solution of (1.1) (see, for instance, Bertsekas [5, Proposition 3.3.2, Exercise 3.7]). The following lemma relates these three conditions with the WCC and SC conditions of Section 2.

LEMMA 3.1.
(i) Condition 2s.3 $\Rightarrow$ Condition 2s.2 $\Rightarrow$ Condition 2s.1.

(ii) *If the SC condition (2.8) holds, then Conditions 2s.2 and 2s.1 are identical.*

(iii) *If the WCC condition (2.9) holds, then Conditions 2s.3 and 2s.2 are identical.*

Proof. The proof of (i) is obvious, since the set of vectors w on which (3.1) is required to hold is successively larger as we go from 2s.1 to 2s.2 to 2s.3. Statement (ii) follows immediately from the definition (2.8) of strict complementarity. For (iii), note that (2.9) implies that

$$\text{null}\,[\nabla g_i(z^*)]^T_{i \in \mathcal{B}_+(\lambda^*)} = \text{null}\,[\nabla g_i(z^*)]^T_{i \in \mathcal{B}_+}, \qquad \text{for all } \lambda^* \in \mathcal{S}_\lambda,$$

from which the result follows immediately. ∎

For any subset $\tilde{\mathcal{B}} \subset \mathcal{B}$ (where $\mathcal{B}$ is the active set defined in (2.6)), we can define the nonlinear program in which just the constraints $i \in \tilde{\mathcal{B}}$ are enforced as follows:

$$(3.3) \qquad \text{NLP}(\tilde{\mathcal{B}}): \qquad\qquad \min_z \phi(z) \qquad \text{subject to } g_i(z) \le 0, \text{ all } i \in \tilde{\mathcal{B}}.$$

Note that any first-order point $(z^*, \lambda^*_{\tilde{\mathcal{B}}})$ for (3.3) can be extended to a first-order point for (1.1) by simply adding zeros to fill out the components $i \notin \tilde{\mathcal{B}}$. Conversely, any solution $(z^*, \lambda^*) \in \mathcal{S}$ for which $\lambda^*_i = 0$, $i \notin \tilde{\mathcal{B}}$, yields a point $(z^*, \lambda^*_{\tilde{\mathcal{B}}})$ that satisfies the first-order conditions for (3.3), by deleting the (zero) components $i \notin \tilde{\mathcal{B}}$ from λ^*. Note, however, that the vector so obtained does not necessarily satisfy second-order conditions for (3.3); in fact, z^* may not even be a local solution for (3.3).

We now define two sets Φ and $\bar{\Phi}$ made up of subsets $\tilde{\mathcal{B}} \subset \mathcal{B}$ as follows:

$$(3.4) \quad \Phi \overset{\text{def}}{=} \{\tilde{\mathcal{B}} \subset \mathcal{B} \,|\, z^* \text{ is a local solution of NLP}(\tilde{\mathcal{B}})$$
$$\text{that satisfies Condition 2s.1 applied to NLP}(\tilde{\mathcal{B}})\},$$

$$(3.5) \quad \bar{\Phi} \overset{\text{def}}{=} \{\tilde{\mathcal{B}} \in \Phi \,|\, \text{the optimal Lagrange multiplier for NLP}(\tilde{\mathcal{B}}) \text{ is unique}\}.$$

When $\tilde{\mathcal{B}} \in \bar{\Phi}$, we use $\lambda^*(\tilde{\mathcal{B}})$ to denote the unique optimal multiplier for $\text{NLP}(\tilde{\mathcal{B}})$, padded out with zeros to length m. Note that $\mathcal{B} \in \Phi$, so that $\Phi \ne \emptyset$. When the strict complementarity and nondegeneracy conditions hold at the solution of (1.1), we have $\Phi = \bar{\Phi} = \{\mathcal{B}\}$.

The sets Φ and $\bar{\Phi}$ become particularly relevant in Section 8, where we propose an algorithm whose steps are obtained by applying SQP to problems of the form $\text{NLP}(\tilde{\mathcal{B}})$. For now, we prove two simple results about the way that these sets are related to each other and to the second-order sufficient conditions.

LEMMA 3.2. *Given some set $\tilde{\mathcal{B}} \in \Phi$, a sufficient condition for $\tilde{\mathcal{B}} \in \bar{\Phi}$ is that the vectors $\{\nabla g_i(z^*), \, i \in \tilde{\mathcal{B}} \cap \mathcal{B}_+\}$, are linearly independent.*

Proof. Since $\tilde{\mathcal{B}} \in \Phi$, there is a vector $\lambda^*_{\tilde{\mathcal{B}}}$ such that

$$(3.6) \qquad\qquad \sum_{i \in \tilde{\mathcal{B}}} \lambda^*_i \nabla g_i(z^*) + \nabla \phi(z^*) = 0.$$

For the components $i \in \tilde{\mathcal{B}} \cap \mathcal{B}_0$, we must have that $\lambda^*_i = 0$, since otherwise the vector $\lambda^*_{\tilde{\mathcal{B}}}$ could be padded out with zeros to yield a vector $\lambda^* \in \mathcal{S}_\lambda$ with $\lambda^*_i \ne 0$ for $i \in \mathcal{B}_0$, contradicting the definition of $\mathcal{B}_+$. Hence we can rewrite (3.6) as

$$\sum_{i \in \tilde{\mathcal{B}} \cap \mathcal{B}_+} \lambda^*_i \nabla g_i(z^*) + \nabla \phi(z^*) = 0,$$

and so linear independence of the given vector set implies uniqueness of $\lambda^*_{\tilde{\mathcal{B}}}$. Therefore $\tilde{\mathcal{B}} \in \bar{\Phi}$, as claimed. ∎

LEMMA 3.3. *Suppose that Condition 2s.3 holds. Then Condition 2s.3 also holds for NLP($\tilde{\mathcal{B}}$), where $\tilde{\mathcal{B}} \subset \mathcal{B}$ is such that there exists a $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subset \tilde{\mathcal{B}}$. In particular, $\tilde{\mathcal{B}} \in \Phi$.*

Proof. Suppose that Condition 2s.3 holds for NLP. For all $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subset \tilde{\mathcal{B}}$, we have by the correspondence between optimal Lagrange multipliers for NLP and NLP($\tilde{\mathcal{B}}$) discussed above that (3.1) holds for all w satisfying (3.2). Hence, Condition 2s.3 holds for the problem NLP($\tilde{\mathcal{B}}$) as well. Since there is at least one vector λ^* with the properties indicated, and since Condition 2s.3 implies Condition 2s.1 by Lemma 3.1(i), we have $\tilde{\mathcal{B}} \in \Phi$. $\square$

4. The iSQP Framework. In the best-known form of the SQP algorithm, the following inequality constrained subproblem is solved to obtain the step at each iteration:

$$(4.1) \qquad \min_{\Delta z} \Delta z^T \nabla \phi(z) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z, \lambda) \Delta z,$$
$$\text{subject to } g(z) + \nabla g(z)^T \Delta z \leq 0,$$

where (z, λ) is the current primal-dual iterate. Denoting the Lagrange multipliers for the constraints in (4.1) by λ^+, we see that the solution Δz satisfies the following KKT conditions (cf. (2.4)):

$$(4.2) \qquad \left[\begin{array}{c} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g(z) \lambda^+ \\ g(z) + \nabla g(z)^T \Delta z \end{array} \right] \in \left[\begin{array}{c} 0 \\ N(\lambda^+) \end{array} \right].$$

We focus our attention, however, on a more general framework that allows for inexactness in the subproblem solution by introducing perturbations into both the objective and constraints of (4.1). We assume only that $(\Delta z, \lambda^+)$ is the solution of

$$(4.3) \qquad \min_{\Delta z} \Delta z^T (\nabla \phi(z) + t) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z, \lambda) \Delta z,$$
$$\text{subject to } g(z) + \nabla g(z)^T \Delta z + r \leq 0,$$

for some perturbation vectors t and r. The KKT conditions for (4.3) become

$$(4.4) \qquad \left[\begin{array}{c} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + t + \nabla g(z) \lambda^+ \\ g(z) + \nabla g(z)^T \Delta z + r \end{array} \right] \in \left[\begin{array}{c} 0 \\ N(\lambda^+) \end{array} \right].$$

We introduce further terminology and notation here: Given a primal-dual solution $(\Delta z, \lambda^+)$ to (4.1) or (4.3), the *strict working set* is the set of indices i for which λ_i^+ is strictly positive. We denote this set by $\mathcal{B}(z, \lambda)$ in the case of (4.1) and $\mathcal{B}(z, \lambda; t, r)$ in the case of (4.3).

When (z, λ) is sufficiently close to $\mathcal{S}$, it happens under mild assumptions that the strict working set $\mathcal{B}(z, \lambda; t, r)$ from an iteration of iSQP identifies a superset of $\mathcal{B}_+(\lambda^*)$ for some optimal multiplier λ^*. This result is interesting for its own sake and also in the context of Section 8, so we prove it here.

LEMMA 4.1. *Suppose that Assumption 1, Condition 2s.1, and MFCQ hold. Then there is a threshold value $\bar{\delta}$ such that whenever $\delta(z, \lambda) \leq \bar{\delta}$ and $\|(t, r)\| \leq \bar{\delta}$, the solution of the iSQP subproblem (4.3) yields a strict working set $\mathcal{B}(z, \lambda; t, r)$ such that $\mathcal{B}_+(\lambda^*) \subset \mathcal{B}(z, \lambda; t, r)$ for at least one $\lambda^* \in \mathcal{S}_\lambda$.*

Proof. Suppose for contradiction that there is a sequence (z^ℓ, λ^ℓ) with $\delta(z^\ell, \lambda^\ell) \downarrow 0$ and a sequence of perturbations (t^ℓ, r^ℓ) with $\|(t^\ell, r^\ell)\| \downarrow 0$ such that the stated property does not hold. That is, taking the active set $\mathcal{B}(z^\ell, \lambda^\ell; t^\ell, r^\ell)$ for the iSQP

subproblem, we find that the subvector $\lambda^*_{\mathcal{B}\backslash\mathcal{B}(z^\ell,\lambda^\ell;t^\ell,r^\ell)}$ is nonzero for all $\lambda^* \in \mathcal{S}_\lambda$. By taking a subsequence if necessary, we can assume that $\mathcal{B}(z^\ell, \lambda^\ell; t^\ell, r^\ell) \equiv \hat{\mathcal{B}} \subset \mathcal{B}$ for all ℓ. By compactness of $\mathcal{S}_\lambda$ (Lemma 2.1) and continuity of $\|\lambda^*_{\mathcal{B}\backslash\hat{\mathcal{B}}}\|$ as a function of λ^*, we have that

$$0 < \zeta \stackrel{\text{def}}{=} \min_{\lambda^* \in \mathcal{S}_\lambda} \left\| \lambda^*_{\mathcal{B}\backslash\hat{\mathcal{B}}} \right\|.$$

From Lemma 5.1, we have that the updated multiplier $(\lambda^\ell)^+$ obtained from the iSQP subproblem (4.3) at $(z^\ell, \lambda^\ell, t^\ell, r^\ell)$ satisfies

$$\delta((\lambda^\ell)^+) = O(\delta(z^\ell)) + O(\|(t^\ell, r^\ell)\|) \downarrow 0.$$

Denoting by $P((\lambda^\ell)^+)$ the projection of $(\lambda^\ell)^+$ onto the set $\mathcal{S}_\lambda$, we have that

$$\delta((\lambda^\ell)^+)^2 = \left\| (\lambda^\ell)^+ - P((\lambda^\ell)^+) \right\|^2 \geq \sum_{i \in \mathcal{B}\backslash\hat{\mathcal{B}}} \left[P((\lambda^\ell)^+)_i \right]^2 \geq \zeta^2 > 0,$$

giving a contradiction. $\square$

5. Local Convergence of iSQP. In this section, we describe the improvement obtained in a single iSQP step (4.3) (alternatively, (4.4)).

First, we apply a result of Robinson [17] to show that small-norm local solutions Δz exist for the inexact subproblem (4.3) provided that (z, λ) is sufficiently close to the solution set $\mathcal{S}$ defined in (2.3b). Our two main results, Theorems 5.2 and 5.3, relate the errors $e_\mathcal{N}(z + \Delta z)$ and $\delta(z + \Delta z, \lambda^+)$ at the new iterate to errors at the current point (z, λ). In particular, Theorem 5.3 demonstrates that superlinear convergence of the primal iterate depends critically on stability of the Lagrange multiplier estimates: λ must not change too much from one iteration to the next.

The first result is obtained by applying Robinson's stability results in [17] to the inexact SQP subproblem (4.3).

LEMMA 5.1. *Suppose that Assumption 1, Condition 2s.1, and MFCQ hold. Then for all (z, λ, t, r) with $\delta(z, \lambda)$ and $\|(t, r)\|$ sufficiently small, the problem (4.3) has a local solution Δz near 0 that satisfies*

(5.1) $$\|\Delta z\| + \delta(\lambda^+) = O(\delta(z)) + O(\|(t, r)\|),$$

where λ^+ is the vector of multipliers corresponding to the solution Δz of (4.3).

Proof. For any fixed $\lambda^* \in \mathcal{S}_\lambda$, consider the problem

(5.2) $$\min_{\Delta z} \Delta z^T \nabla \phi(z^*) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z^*, \lambda^*) \Delta z,$$
          $$\text{subject to } g(z^*) + \nabla g(z^*)^T \Delta z \leq 0,$$

whose primal-dual solution set is $\{0\} \times \mathcal{S}_\lambda$. MFCQ holds for this problem, since the active set $\mathcal{B}$ is the same as for the nonlinear problem (1.1), and it is easy to see that Condition 2s.1 is satisfied as well.

Consider now the parametrized version (4.3) of the problem (5.2), in which the parametrization is defined by the vector $p = (z, \lambda, t, r)$. (We recover (5.2) from (4.3) by setting $p = p_0 = (z^*, \lambda^*, 0, 0)$.) We define the following subsets of R^n:

        $\text{stat}(p) = \{\Delta z \,|\, (\Delta z, \lambda^+) \text{ satisfies KKT conditions for (4.3) for some } \lambda^+\},$
        $\text{lsol}(p) = \{\Delta z \,|\, \Delta z \text{ is a local solution of (4.3)}\}.$

By applying Theorem 3.2 of Robinson [17], we find that there is a neighborhood $N_1(\lambda^*)$ of p_0 and a neighborhood $M_1(\lambda^*)$ of 0 such that

$\quad\quad$ stat$(p) \cap M_1(\lambda^*)$ is continuous in p at p_0;

$\quad\quad$ lsol$(p) \cap M_1(\lambda^*)$ is nonempty for $p \in N_1(\lambda^*)$, and a subset of stat$(p) \cap M_1(\lambda^*)$.

Moreover, we have from Theorem 4.2 of Robinson [17] that there is a neighborhood $N_2(\lambda^*)$ of p_0 and a constant $\epsilon(\lambda^*) > 0$ such that for each $p \in N_2(\lambda^*)$ we have that all stationary pairs $(\Delta z, \lambda^+)$ for (4.3) satisfy

$$\inf_{\bar{\lambda} \in \mathcal{S}_\lambda} \|(\Delta z, \lambda^+ - \bar{\lambda})\|$$

$$\leq \epsilon(\lambda^*) \left\| \begin{bmatrix} (\mathcal{L}_{zz}(z,\lambda) - \mathcal{L}_{zz}(z^*,\lambda^*))\Delta z + (\nabla\phi(z) - \nabla\phi(z^*)) + (\nabla g(z) - \nabla g(z^*))\lambda^+ + t \\ (g(z) - g(z^*)) + (\nabla g(z) - \nabla g(z^*))^T \Delta z + r \end{bmatrix} \right\|$$

For the left-hand side of this expression, it is easy to see that

$$(5.4) \quad\quad \inf_{\bar{\lambda} \in \mathcal{S}_\lambda} \|(\Delta z, \lambda^+ - \bar{\lambda})\| \geq \frac{1}{\sqrt{2}} \left(\|\Delta z\| + \delta(\lambda^+) \right).$$

For the right-hand side, we have by Lipschitz continuity of $\nabla^2 \phi$ and $\nabla^2 g_i$, $i = 1, 2, \ldots, m$ that

$$\|[\mathcal{L}_{zz}(z,\lambda) - \mathcal{L}_{zz}(z^*,\lambda^*)]\Delta z\| \leq C_1 \left(\delta(z) + \|\lambda - \lambda^*\| \right) \|\Delta z\|,$$

$$\|\nabla\phi(z) - \nabla\phi(z^*)\| \leq C_1 \delta(z),$$

$$\|g(z) + \nabla g(z)^T \Delta z - g(z^*) - \nabla g(z^*)^T \Delta z\| \leq C_1 \delta(z) \left(1 + \|\Delta z\| \right),$$

for some constant C_1 (which is, in particular, independent of the multiplier λ^*). Moreover, by boundedness of $\mathcal{S}_\lambda$, we have after a possible adjustment of C_1 that

$$\left\| \nabla g(z)\lambda^+ - \nabla g(z^*)\lambda^+ \right\| \leq C_1 \|\lambda^+\| \delta(z) \leq C_1 \left(C_2 + \delta(\lambda^+) \right) \delta(z),$$

where C_2 is a constant that bounds the norms of all elements of $\mathcal{S}_\lambda$. By using these expressions together with (5.4) in (5.3), we have

$$(5.5) \quad \|\Delta z\| + \delta(\lambda^+)$$
$$\leq \epsilon(\lambda^*) C_3 \left[\|\lambda - \lambda^*\| \|\Delta z\| + \delta(z)\|\Delta z\| + \delta(z) + \delta(\lambda^+)\delta(z) + \|(t,r)\| \right],$$

where C_3 is a constant (in particular, independent of the particular choice of $\lambda^* \in \mathcal{S}_\lambda$). By reducing the size of $N_2(\lambda^*)$ if necessary, we have that

$$(5.6) \quad\quad \epsilon(\lambda^*) C_3 \left[\|\lambda - \lambda^*\| + \delta(z) \right] \leq 1/4 \text{ for all } (z, \lambda, t, r) \in N_2(\lambda^*).$$

Thus, by transferring terms involving $\|\Delta z\|$ and $\delta(\lambda^+)$ from the right-hand side to the left-hand side in (5.5), we obtain

$$\|\Delta z\| \left[1 - \epsilon(\lambda^*) C_3 (\|\lambda - \lambda^*\| + \delta(z)) \right]$$
$$+ \delta(\lambda^+) \left[1 - \epsilon(\lambda^*) C_3 \delta(z) \right] \leq \epsilon(\lambda^*) C_3 \left[\delta(z) + \|(t,r)\| \right],$$

so that from (5.6), we have that

$$(5.7) \quad\quad \|\Delta z\| + \delta(\lambda^+) \leq 2\epsilon(\lambda^*) C_3 \left[\delta(z) + \|(t,r)\| \right].$$

Note that (5.7) holds for the *fixed* choice of λ^* in $\mathcal{S}_\lambda$, and only for (z, λ, t, r) within the neighborhood $N_2(\lambda^*)$ of $(z^*, \lambda^*, 0, 0)$. Since

$$\{N_2(\lambda^*) \,|\, \lambda^* \in \mathcal{S}_\lambda\}$$

is a cover of the set

$$(5.8) \qquad \{z^*\} \times \mathcal{S}_\lambda \times \{0\} \times \{0\},$$

we have by compactness of $\mathcal{S}_\lambda$ (and hence of (5.8)) that there is a finite subcover, say

$$\{N_2(\lambda_l^*) \mid \lambda_l^* \in \mathcal{S}_\lambda, \ l = 1, 2, \dots, L\}.$$

Note that the set

$$\cup_{l=1,2,\dots,L} N_2(\lambda_l^*)$$

is a neighborhood of (5.8).

By setting

$$C_4 = 2C_3 \max_{l=1,2,\dots,L} \epsilon(\lambda_l^*),$$

we have from (5.7) that

$$(5.9) \qquad \|\Delta z\| + \delta(\lambda^+) \leq C_4 \left[\delta(z) + \|(t,r)\|\right],$$

whenever

$$(5.10) \qquad (z, \lambda, t, r) \in \cup_{l=1,2,\dots,L} N_2(\lambda_l^*).$$

We can choose $\bar\delta$ sufficiently small that

$$\delta(z,\lambda) \leq \bar\delta, \|(t,r)\| \leq \bar\delta \ \Rightarrow \ (z,\lambda,t,r) \in \cup_{l=1,2,\dots,L} N_2(\lambda_l^*),$$

so that (5.9) holds for all (z, λ) with $\delta(z,\lambda) \leq \bar\delta$ and $\|(t,r)\| \leq \bar\delta$. $\square$

In subsequent discussions, we use the term "iSQP" to describe the inexact SQP procedure in which each iteration consists of obtaining a solution to the problem (4.3) from the current iterate (z, λ) and then setting

$$(5.11) \qquad (z, \lambda) \leftarrow (z + \Delta z, \lambda^+),$$

where $(\Delta z, \lambda^+)$ is a primal-dual solution of (4.3) that satisfies (5.1).

The next result shows that while the iSQP step may not give a "superlinear" decrease in distance to the solution set, it does reduce the error substantially in the $e_{\mathcal{N}}(\cdot)$ component of the primal error vector. (This result explains an observation made while doing computational experiments for an earlier paper [18]. It is similar to Lemma 3.12 of Fischer [8], though the latter result assumes WCC (2.9), which is not needed below.)

THEOREM 5.2. *Suppose that Assumption 1, Condition 2s.1, and MFCQ hold. For all (z, λ) with $\delta(z, \lambda)$ and $\|(t,r)\|$ sufficiently small, the new iterate generated by the iSQP algorithm satisfies*

$$(5.12) \qquad \|e_{\mathcal{N}}(z + \Delta z)\| = O(\delta(z)^2) + O(\|t\|^2) + O(\|r\|).$$

Moreover, we have

$$(5.13) \qquad g_{\mathcal{B}_+}(z + \Delta z) = O(\delta(z)^2) + O(\|t\|^2) + O(\|r\|).$$

Proof. Let $\{(z^k, \lambda^k, t^k, r^k)\}$ be any sequence with

$$\delta(z^k, \lambda^k) \to 0, \qquad \|(t^k, r^k)\| \to 0,$$

and let $(\Delta z^k, \lambda^{k+})$ be the primal-dual solution to (4.3) obtained when $(z, \lambda) = (z^k, \lambda^k)$ and $(t, r) = (t^k, r^k)$. Denote the corresponding sequence of strict working sets for the iSQP subproblem (4.3) by $\mathcal{B}^k$. From Lemma 5.1, we have $\Delta z^k \to 0$, so since $g(z^k) \to g(z^*)$ and $r^k \to 0$, none of the inactive indices $j \notin \mathcal{B}$ can be active in the subproblem for k sufficiently large. We can therefore assume without loss of generality that $\mathcal{B}^k \subset \mathcal{B}$.

Note first that from (5.1) and $\|e(z)\| = \delta(z)$, we have

$$(5.14) \qquad \|e(z^k + \Delta z^k)\| \leq \|e(z^k)\| + \|\Delta z^k\| = O(\delta(z^k)) + O(\|(t^k, r^k)\|).$$

For all active indices $i \in \mathcal{B}^k$, we have from Taylor's theorem together with (5.14) and (4.3) that

$$
\begin{aligned}
\nabla g_i(z^*)^T e(z^k + \Delta z^k) &= g_i(z^k + \Delta z^k) - g_i(z^*) + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&= g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&= -r_i^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
(5.15) \qquad &= \tilde{r}_i^k,
\end{aligned}
$$

where

$$(5.16) \qquad \tilde{r}_i^k = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|).$$

Meanwhile, for $i \in \mathcal{B} \backslash \mathcal{B}^k$, we have from (4.2) and (5.1) that

$$
\begin{aligned}
\nabla g_i(z^*)^T e(z^k + \Delta z^k) &= g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
&\leq -r_i^k + O(\delta(z^k)^2) + O(\|(t^k, r^k)\|^2) \\
(5.17) \qquad &\leq \tilde{r}_i^k,
\end{aligned}
$$

where the estimate (5.16) holds once again.

Since $(\Delta z^k, \lambda^{k+})$ is the solution to (4.3), then by boundedness of $\mathcal{S}_\lambda$, we have from (4.4) and (5.1) that

$$\nabla \phi(z^k) + \mathcal{L}_{zz}(z^k, \lambda^k)\Delta z^k + \sum_{i \in \mathcal{B}^k} \lambda_i^{k+} \nabla g_i(z^k) + t^k = 0$$

$$\Rightarrow \nabla \phi(z^*) + \sum_{i \in \mathcal{B}^k} \lambda_i^{k+} \nabla g_i(z^*) = O(\delta(z^k)) + O(\|(t^k, r^k)\|).$$

By the definition (2.7) of $\mathcal{B}_+$, there is a $\lambda^* \in \mathcal{S}_\lambda$ such that

$$(5.18) \qquad \nabla \phi(z^*) + \sum_{i \in \mathcal{B}_+} \lambda_i^* \nabla g_i(z^*) = 0, \qquad \lambda_{\mathcal{B}_+}^* > 0.$$

(We can construct λ^* by taking $\lambda^{(i)} \in \mathcal{S}_\lambda$ with $\lambda_i^{(i)} > 0$ for each $i \in \mathcal{B}_+$, according to the definition (2.7), and setting $\lambda^* = \sum_{i \in \mathcal{B}_+} \lambda^{(i)}/|\mathcal{B}_+|$.) By combining the last two equations, we obtain

$$\sum_{i \in \mathcal{B}_+ \backslash \mathcal{B}^k} \lambda_i^* \nabla g_i(z^*) = \sum_{i \in \mathcal{B}^k} (\lambda_i^{k+} - \lambda_i^*) \nabla g_i(z^*) + O(\delta(z^k)) + O(\|(t^k, r^k)\|).$$

By taking inner products of both sides with $e(z^k + \Delta z^k)$, and using (5.14), we have by (5.15), (5.16), and boundedness of $\mathcal{S}_\lambda$ that

$$\sum_{i \in \mathcal{B}_+ \backslash \mathcal{B}^k} \lambda_i^* \nabla g_i(z^*)^T e(z^k + \Delta z^k) = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|).$$

Since $\lambda_{\mathcal{B}_+}^* > 0$, and since by (5.17) none of the terms $\nabla g_i(z^*)^T e(z^k + \Delta z^k)$, $i \in \mathcal{B}_+ \backslash \mathcal{B}^k$ can be larger than a small positive number of the size indicated in (5.16), we have that

$$\nabla g_i(z^*)^T e(z^k + \Delta z^k) = \tilde{r}_i^k \quad \text{for all } i \in \mathcal{B}_+ \backslash \mathcal{B}^k,$$

where the values of $\tilde{r}_i^k$ may have been adjusted from (5.17) but they still satisfy the estimate (5.16). Hence, for the indices $i \in \mathcal{B}_+ \backslash \mathcal{B}^k$, we can replace the inequality by an equality in (5.17). By combining this observation with (5.15) and (5.17), we find that

$$\text{(5.19a)} \qquad \nabla g_{\mathcal{B}_+}(z^*)^T e(z^k + \Delta z^k) = \tilde{r}_{\mathcal{B}_+}^k,$$

$$\text{(5.19b)} \qquad \nabla g_{\mathcal{B}_0}(z^*)^T e(z^k + \Delta z^k) \le \tilde{r}_{\mathcal{B}_0}^k,$$

where

$$\text{(5.20)} \qquad \|\tilde{r}_{\mathcal{B}}^k\| = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|).$$

Consider now the partitioning of $e(z^k + \Delta z^k)$ into its $e_{\mathcal{N}}(\cdot)$ and $e_{\mathcal{T}}(\cdot)$ components. From (2.13), we see that the $e_{\mathcal{T}}$ component is obtained by solving

$$\text{(5.21)} \qquad \min_{e_{\mathcal{T}}} \frac{1}{2} \|e_{\mathcal{T}} - e(z^k + \Delta z^k)\|^2,$$
$$\text{subject to } \nabla g_{\mathcal{B}_+}(z^*)^T e_{\mathcal{T}} = 0, \ \nabla g_{\mathcal{B}_0}(z^*)^T e_{\mathcal{T}} \le 0.$$

The problem (5.21) is a feasible and strictly convex problem, so it has a unique solution. Consider too the following perturbation:

$$\text{(5.22)} \qquad \min_{e_{\mathcal{T}}} \frac{1}{2} \|e_{\mathcal{T}} - e(z^k + \Delta z^k)\|^2,$$
$$\text{subject to } \nabla g_{\mathcal{B}_+}(z^*)^T e_{\mathcal{T}} = \tilde{r}_{\mathcal{B}_+}^k, \ \ \nabla g_{\mathcal{B}_0}(z^*)^T e_{\mathcal{T}} \le \tilde{r}_{\mathcal{B}_0}^k,$$

for which the (unique) solution is $e(z^k + \Delta z^k)$, because of (5.19). By applying Lemma B.1, we have that the solutions of (5.21) and (5.22) are related as follows:

$$\|e_{\mathcal{T}}(z^k + \Delta z^k) - e(z^k + \Delta z^k)\| = O(\|\tilde{r}_{\mathcal{B}}^k\|) = O(\delta(z^k)^2) + O(\|t^k\|^2) + O(\|r^k\|).$$

Since $e_{\mathcal{N}}(\cdot) = e(\cdot) - e_{\mathcal{T}}(\cdot)$, the result (5.12) follows immediately.

The second part of the result follows readily from a Taylor series argument, together with (5.19a), (5.14), and the estimate of $\|\tilde{r}^k\|$. $\square$

We are now ready to prove the result about local convergence of the iSQP algorithm.

THEOREM 5.3. *Suppose that Assumption 1, Condition 2s.1, and the MFCQ condition hold. Suppose that a new iterate $(z+\Delta z, \lambda^+)$ is generated by the iSQP algorithm*

from the point (z, λ). Then for all (z, λ, t, r) with $\delta(z, \lambda)$ and $\|(t, r)\|$ sufficiently small, we have

$$(5.23) \qquad \delta(z + \Delta z, \lambda^+) = \|\lambda - \lambda^+\| O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r)\|).$$

For the special case in which $g(\cdot)$ is linear, we have

$$(5.24) \qquad \delta(z + \Delta z, \lambda^+) = O(\delta(z)^2) + O(\|(t, r)\|).$$

Proof. Consider the problem

$$(5.25) \qquad \begin{bmatrix} \nabla\phi(\tilde{z}) + \nabla g(\tilde{z})\tilde{\lambda} + \tilde{t} \\ g(\tilde{z}) + \tilde{r} \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\tilde{\lambda}) \end{bmatrix},$$

where

$$(5.26a) \qquad \tilde{t} = \mathcal{L}_{zz}(z, \lambda)\Delta z + \nabla\phi(z) + t + \nabla g(z)\lambda^+$$
$$-\nabla\phi(z + \Delta z) - \nabla g(z + \Delta z)\lambda^+,$$
$$(5.26b) \qquad \tilde{r} = g(z) + \nabla g(z)^T \Delta z + r - g(z + \Delta z).$$

By (4.4), the solution of (5.25) is simply $(\tilde{z}, \tilde{\lambda}) = (z + \Delta z, \lambda^+)$. Viewing (5.25) as a perturbed version of (2.4), we can apply Corollary 4.3 of Robinson [17] to deduce that

$$(5.27) \qquad \delta(z + \Delta z, \lambda^+) = O(\|(\tilde{t}, \tilde{r})\|).$$

By using the assumed smoothness of ϕ and g, we have

$$\tilde{t} = \nabla^2\phi(z)\Delta z + \sum_{i=1}^m \lambda_i \nabla^2 g_i(z)\Delta z - [\nabla\phi(z + \Delta z) - \nabla\phi(z)]$$
$$-[\nabla g(z + \Delta z) - \nabla g(z)]\lambda^+ + t$$
$$= \sum_{i=1}^m (\lambda_i - \lambda_i^+)\nabla^2 g_i(z)\Delta z + O(\|\Delta z\|^2) + t$$

so by using boundedness of the sets containing λ and λ^+ and the estimate (5.1), we obtain

$$\|\tilde{t}\| \le \|\lambda - \lambda^+\| [O(\delta(z)) + O(\|(t, r)\|)] + O(\delta(z)^2) + O(\|(t, r)\|^2) + O(\|t\|)$$
$$(5.28) \qquad = \|\lambda - \lambda^+\| O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r)\|).$$

For $\tilde{r}$, we have

$$(5.29) \qquad \tilde{r} = O(\|\Delta z\|^2) + r = O(\delta(z)^2) + O(\|(t, r)\|).$$

The result (5.23) is immediate from (5.27), (5.28), and (5.29).

The second result (5.24) also is immediate if we observe that the term containing $\lambda - \lambda^+$ vanishes from $\tilde{t}$ when $g(\cdot)$ is linear. $\square$

6. Stabilized SQP. Superlinear convergence of the stabilized SQP method (sSQP) has been discussed by by Wright [18] and Hager [11]. We show here that this method can be placed in the iSQP framework and that the superlinear convergence result therefore can be derived from Theorem 5.3.

The sSQP algorithm is derived by applying proximal point ideas to the SQP subproblem (4.1). Specifically, it adds a term to the objective for (4.1) that penalizes the step from λ to λ^+. From a current primal-dual iterate (z, λ), we find a local solution of the following minimax subproblem for $(\Delta z, \lambda^+)$ such that $(\Delta z, \lambda^+ - \lambda)$ is small:

$$(6.1) \qquad \min_{\Delta z} \max_{\lambda^+ \geq 0} \Delta z^T \nabla \phi(z) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z, \lambda) \Delta z$$
$$+ (\lambda^+)^T [g(z) + \nabla g(z)^T \Delta z] - \tfrac{1}{2} \mu \|\lambda^+ - \lambda\|^2,$$

where μ is a positive parameter that may be varied from one iteration to the next. The first-order conditions for $(\Delta z, \lambda^+)$ to solve (6.1) are

$$(6.2) \qquad \begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g(z) \lambda^+ \\ g(z) + \nabla g(z)^T \Delta z - \mu(\lambda^+ - \lambda) \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+) \end{bmatrix}.$$

It is easy to show that for $\delta(z, \lambda)$ sufficiently small, any solution to (6.2) with $\|\Delta z\|$ small must have $\lambda_i^+ = 0$ for $i \notin \mathcal{B}$. For such indices i, we have

$$g_i(z) + \nabla g_i(z)^T \Delta z - \mu(\lambda_i^+ - \lambda_i) \leq g_i(z) + \nabla g_i(z)^T \Delta z + \mu \lambda_i \leq (1/2) g_i(z^*) < 0,$$

when the second inequality holds whenever $\delta(z, \lambda)$ and $\|\Delta z\|$ are sufficiently small. By complementarity, it follows that $\lambda_i^+ = 0$, as claimed. Therefore we can asymptotically drop the inactive constraints from consideration. Denoting by $\tilde{\mathcal{B}} \subset \mathcal{B}$ the subset of active indices in (6.2) (so that $\lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}}^+ = 0$), we have by partitioning indices that

$$\begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g_{\tilde{\mathcal{B}}}(z) \lambda_{\tilde{\mathcal{B}}}^+ \\ g_{\tilde{\mathcal{B}}}(z) + \nabla g_{\tilde{\mathcal{B}}}(z)^T \Delta z - \mu(\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}) \\ g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z) + \nabla g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z)^T \Delta z + \mu \lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}} \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda_{\tilde{\mathcal{B}}}^+) \\ N(\lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}}^+) \end{bmatrix}.$$

Since $\mu \lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}} \geq 0$, the pair $(\Delta z, \lambda^+)$ that solves this system also satisfies

$$(6.3) \qquad \begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) \Delta z + \nabla \phi(z) + \nabla g_{\tilde{\mathcal{B}}}(z) \lambda_{\tilde{\mathcal{B}}}^+ \\ g_{\tilde{\mathcal{B}}}(z) + \nabla g_{\tilde{\mathcal{B}}}(z)^T \Delta z - \mu(\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}) \\ g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z) + \nabla g_{\mathcal{B} \setminus \tilde{\mathcal{B}}}(z)^T \Delta z \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda_{\tilde{\mathcal{B}}}^+) \\ N(\lambda_{\mathcal{B} \setminus \tilde{\mathcal{B}}}^+) \end{bmatrix}.$$

Hence, we can view sSQP as a special case of (4.4) in which we have

$$(6.4) \qquad t = 0, \quad r_{\tilde{\mathcal{B}}} = -\mu(\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}), \quad r_{\{1,\ldots,m\} \setminus \tilde{\mathcal{B}}} = 0.$$

There is no circular logic here in the choice of λ^+. If we fix λ^+ at its optimal value from (6.1) and fix t and r in (4.4) at the values in (6.4), then the same $(\Delta z, \lambda^+)$ that solves (6.1) (and (6.3)) will solve (4.4).

A form of the sSQP algorithm was proposed earlier by Bartholomew-Biggs [3]. The basic steps generated by Bartholomew-Biggs' algorithm have the form indicated above (except that a quasi-Newton approximation is used in place of the actual Hessian of the Lagrangian). However, there are numerous modifications that place the

algorithm in the global convergence framework of that author's REQP algorithm [1, 2]. For instance, the multiplier estimates λ are not necessarily updated at each iteration, and successive values of μ are chosen by heuristics rather than by an estimate of the distance to the solution set, as is the case in [18]. The focus of Bartholomew-Biggs' work is somewhat complementary to that of Wright and Hager, since the latter concerns itself with local convergence issues (for the case in which MFCQ is satisfied), while the former focuses on global convergence. The sSQP approach is also closely related to a variant of the method of multipliers/augmented Lagrangian algorithm (see Bertsekas [4] and [5, Section 4.2]) in which just one Newton step is applied to the augmented Lagrangian function at each iteration, and in which the parameter μ is decreased to zero. Indeed, such a variant is given for the case of equality constrained problems by Bertsekas [4, p. 240], who points out its superlinear local convergence properties (for the case in which the active constraint gradients are linearly independent).

Superlinear convergence of the sSQP algorithm can be proved if the stabilization parameter μ is related appropriately to the distance $\delta(z, \lambda)$ from (z, λ) to the solution set $\mathcal{S}$. Such an estimate is readily available; we show in the Appendix (Theorem A.1) that

$$(6.5) \qquad \eta(z, \lambda) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \mathcal{L}_z(z, \lambda) \\ \min(\lambda, -g(z)) \end{bmatrix} \right\| = \Theta(\delta(z, \lambda)),$$

where the "min" operation applies componentwise to the argument vectors.

Suppose now that we choose μ to satisfy

$$(6.6) \qquad \mu = \eta(z, \lambda)^\tau,$$

where $\tau \in (0, 1)$. From (6.3), since $\tilde{\mathcal{B}}$ denotes the active constraints in (6.2), we have that

$$(6.7) \qquad \begin{bmatrix} \mathcal{L}_{zz}(z, \lambda) & \nabla g_{\tilde{\mathcal{B}}}(z) \\ \nabla g_{\tilde{\mathcal{B}}}(z)^T & -\mu I \end{bmatrix} \begin{bmatrix} \Delta z \\ \lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}} \end{bmatrix} = \begin{bmatrix} -\nabla \phi(z) - \nabla g_{\tilde{\mathcal{B}}}(z) \lambda_{\tilde{\mathcal{B}}} \\ -g_{\tilde{\mathcal{B}}}(z) \end{bmatrix},$$

Hager's proof of [11, Theorem 1] shows that the subset $\tilde{\mathcal{B}} \subset \mathcal{B}$ has the property that

z^* is a local minimum of $\text{NLP}(\tilde{\mathcal{B}})$ that satisfies Condition 2s.3 applied to $\text{NLP}(\tilde{\mathcal{B}})$,

while $\lambda_i^+ = 0$, $i \notin \tilde{\mathcal{B}}$. From [11, Theorem 1], we have that

$$(6.8) \qquad \|\lambda^+ - \lambda\| = O(\delta(z, \lambda)).$$

Therefore if we define (t, r) as in (6.4), we have from (6.5), (6.6), and (6.8) that

$$(6.9) \qquad (t, r) = O(\delta(z, \lambda)^{1+\tau}).$$

By substituting (6.8) and (6.9) into (5.23), we obtain

$$\delta(z + \Delta z, \lambda^+) = O(\delta(z, \lambda)) O(\delta(z)) + O(\delta(z, \lambda)^{1+\tau}) = O(\delta(z, \lambda)^{1+\tau}).$$

Hence, the convergence rate of sSQP can be derived by placing it in the framework of iSQP.

The same result can be derived from the results of Hager [11]. The following result is a simple consequence of [11, Theorem 1].

THEOREM 6.1. *Suppose that Assumption 1, Condition 2s.3, and the MFCQ condition hold. Suppose, too, that μ defined by (6.6) is used as the stabilization parameter at each iteration of sSQP. Then there exists a positive threshold $\bar{\delta}$ such that for any (z, λ) with $\delta(z, \lambda) \leq \bar{\delta}$, there exists a local solution $(\Delta z, \lambda^+)$ of the sSQP subproblem (6.1) such that*

$$(6.10) \qquad\qquad \delta(z + \Delta z, \lambda^+) = O(\delta(z, \lambda)^{1+\tau}).$$

Proof. From Condition 2s.3, we have for each $\lambda^* \in \mathcal{S}_\lambda$ that $w^T \mathcal{L}_{zz}(z^*, \lambda^*) w \geq \sigma \|w\|^2$ for some $\sigma > 0$ and all w with $\nabla g_{\mathcal{B}_+(\lambda^*)}(z^*)^T w = 0$. Moreover the choice (6.6) ensures that we have

$$\sigma_0 \delta(z, \lambda) \leq \mu \leq \sigma_1,$$

for $\delta(z, \lambda)$ sufficiently small, where σ_0 and σ_1 are constants defined in Theorem 1 of Hager [11], with σ_0 "sufficiently large." By applying Hager's result, we find that there is a neighborhood $\mathcal{U}(\lambda^*)$ of (z^*, λ^*) such that if $(z, \lambda) \in \mathcal{U}(\lambda^*)$, then the sSQP subproblem (6.1) yields a solution in $\mathcal{U}(\lambda^*)$ such that (6.10) is satisfied. Note that the set

$$\{\mathcal{U}(\lambda^*) \,|\, \lambda^* \in \mathcal{S}_\lambda\}$$

forms a cover of $\mathcal{S}$. By compactness, we can select a finite subcover

$$\{\mathcal{U}(\lambda_1^*), \mathcal{U}(\lambda_2^*), \ldots, \mathcal{U}(\lambda_p^*)\},$$

for some $\lambda_1^*, \lambda_2^*, \ldots, \lambda_p^* \in \mathcal{S}_\lambda$. By choosing $\bar{\delta}$ positive but small enough that

$$\delta(z, \lambda) \leq \bar{\delta} \quad \Rightarrow \quad (z, \lambda) \in \mathcal{U}(\lambda_1^*) \cup \mathcal{U}(\lambda_2^*) \cup \cdots \cup \mathcal{U}(\lambda_p^*),$$

we obtain the desired result. ∎

In Wright [18], it was shown that if the initial estimate λ^0 is not too close to the boundary of $\mathcal{S}_\lambda$ (in the sense that $\lambda_i^0 \geq \xi$ for some $\xi > 0$ and all $i \in \mathcal{B}$), then all steps are obtained from a system of the form (6.7) with $\tilde{\mathcal{B}} = \mathcal{B}$. Moreover, we can set $\tau = 1$ in (6.6) (yielding a quadratic rate of convergence), and we need assume only that the weaker Condition 2s.1 holds. Implementation of such an approach would not be difficult, since it requires only a reliable way to estimate the active set $\mathcal{B}$, along with solution of a subproblem to adjust $\lambda_{\mathcal{B}}$ so that all components of this vector are sufficiently positive.

7. Fischer's Method. Fischer's method, as described in the paper [8], generates steps Δz in the primal variables by solving a standard SQP subproblem. The Lagrange multiplier estimate obtained from this subproblem is then discarded, and an auxiliary subproblem of similar complexity to the SQP subproblem is then solved to obtain the new multiplier estimate. Superlinear convergence of the resulting algorithm is proved in [8], under assumptions that we discuss later.

Fischer's method can be described in terms of the iSQP framework of Section 4 as analyzed in Section 5. We can show that the primal step Δz generated by this method can be embedded in a primal-dual solution $(\Delta z, \tilde{\lambda}^+)$ to an *inexact* SQP subproblem of the form (4.3), so that Theorem 5.3 applies. Superlinear convergence of the primal iterates then follows from the fact that the difference between $\tilde{\lambda}^+$ and Fischer's specific

Lagrange multiplier estimate $\hat{\lambda}$ has magnitude $O(\delta(z))$. Superlinear convergence of Fischer's Lagrange multiplier estimates $\hat{\lambda}$ to the set $\mathcal{S}_\lambda$ follows from the fact that $\delta(\hat{\lambda}) = O(\delta(z))$.

A single step of Fischer's algorithm proceeds as follows. First, given the primal iterate z, the following subproblem is solved to find the pair $(d, \hat{\lambda})$:

$$(7.1) \qquad \begin{bmatrix} \nabla\phi(z) + d + \nabla g(z)\hat{\lambda} \\ \nabla g(z)^T d + g(z) \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\hat{\lambda}) \end{bmatrix}.$$

The primal component d is discarded, and $\hat{\lambda}$ is adopted as the new dual iterate. The next primal step is then obtained by solving the SQP subproblem (4.2) from the point $(z, \hat{\lambda})$, that is,

$$(7.2) \qquad \begin{bmatrix} \mathcal{L}_{zz}(z,\hat{\lambda})\Delta z + \nabla\phi(z) + \nabla g(z)\lambda^+ \\ g(z) + \nabla g(z)^T \Delta z \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda^+) \end{bmatrix}.$$

The dual component λ^+ is now discarded (indeed, there is no need to calculate it at all), and this iteration is complete. The next iteration begins by solving (7.1) again, with $z + \Delta z$ replacing z, to obtain the new multiplier estimate $\hat{\lambda}^+$.

Note that in the auxiliary problem (7.1), $(d, \hat{\lambda})$ is the primal-dual solution of the problem

$$\min_d \frac{1}{2}d^T d + d^T \nabla\phi(z), \quad \text{subject to } \nabla g(z)^T d + g(z) \leq 0$$

(see Fischer [8, p. 13]). More tellingly, we can view (7.1) as a perturbation of the problem

$$(7.3) \qquad \begin{bmatrix} \nabla\phi(z^*) + d + \nabla g(z^*)\hat{\lambda} \\ \nabla g(z^*)^T d + g(z^*) \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\hat{\lambda}) \end{bmatrix},$$

in which z has been replaced by z^*. Noting that the solution set for (7.3) is $(d, \hat{\lambda}) \in 0 \times \mathcal{S}_\lambda$, we can again apply Robinson's results from [17] (and in particular [17, Corollary 4.3]) to obtain the estimate

$$(7.4) \qquad \|d\| + \delta(\hat{\lambda}) = O(\delta(z)),$$

for all solutions $(d, \hat{\lambda})$ of (7.1).

Fischer [8, Theorem 3.13] shows that under certain assumptions (discussed below), the primal component Δz for the solution of (7.2) is also the solution of the following iSQP subproblem:

$$(7.5) \qquad \begin{bmatrix} \mathcal{L}_{zz}(z,\hat{\lambda})\Delta z + \nabla\phi(z) + \nabla g(z)\tilde{\lambda}^+ + t \\ g(z) + \nabla g(z)^T \Delta z + r \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\tilde{\lambda}^+) \end{bmatrix},$$

where the perturbation vectors t and r and the multiplier estimates $\tilde{\lambda}^+$ satisfy

$$(7.6) \qquad t = 0, \qquad \|r\| = O(\delta(z)^2), \qquad \|\tilde{\lambda}^+ - \hat{\lambda}\| = O(\delta(z)).$$

Hence, Theorem 5.3 can be applied to deduce that

$$(7.7) \quad \delta(z + \Delta z, \tilde{\lambda}^+) = \|\tilde{\lambda}^+ - \hat{\lambda}\|O(\delta(z)) + O(\delta(z)^2) + O(\|(t,r)\|) = O(\delta(z)^2).$$

Since $\delta(z + \Delta z) \le \delta(z + \Delta z, \tilde{\lambda}^+)$, this expression implies Q-quadratic convergence in the primal iterates. Q-quadratic convergence of the primal-dual iterates follows from (7.4). Note that the multipliers $\tilde{\lambda}^+$ are never calculated explicitly by the algorithm.

The assumptions needed to prove Theorem 3.13 in [8] include the WCC condition (2.9), the MFCQ condition (2.10), the second-order sufficient Condition 2s.1, and the following constant-rank condition:

$$(7.8) \qquad\qquad \nabla g_{\mathcal{B}_+}(z) \text{ has constant rank for all } z \text{ near } z^*.$$

The need to solve the auxiliary subproblem (7.1) at every iteration is a disadvantage of Fischer's approach.

8. SQP with Strict Working Sets. Although the preceding two sections show that modified versions of SQP algorithms converge superlinearly on degenerate problems (under certain assumptions), practical experience shows that standard SQP strategies usually encounter little trouble with problems of this type. Frequently, the strict working sets $\mathcal{B}^k$ for the QP subproblems settle down to a constant set in the neighborhood of the solution, and the Lagrange multiplier estimates often approach a unique limit.

As we saw in Theorem 5.3, superlinear convergence depends critically on stabilization of the Lagrange multiplier estimates λ^k. Such stabilization is guaranteed to occur if the strict working sets $\mathcal{B}^k$ eventually become subsets of some fixed set set $\tilde{\mathcal{B}} \subset \mathcal{B}$ with the following properties:

(i)   there is a unique multiplier $\lambda^*(\tilde{\mathcal{B}}) \in \mathcal{S}_\lambda$ such that $\lambda_i^*(\tilde{\mathcal{B}}) = 0$ for $i \notin \tilde{\mathcal{B}}$; and

(ii) the nonlinear program $\mathrm{NLP}(\tilde{\mathcal{B}})$ obtained by dropping the constraints $i \notin \tilde{\mathcal{B}}$ from (1.1) still has a minimizer at z^* that satisfies Condition 2s.1.

If these properties hold, the only possible limit for the sequence of Lagrange multiplier estimates λ^k is the unique vector $\lambda^*(\tilde{\mathcal{B}})$ defined in (i). Recall from the definition (3.5) that $\bar{\Phi}$ contains precisely those subsets of $\mathcal{B}$ with properties (i) and (ii).

The code SNOPT [10] is a recent implementation of SQP that appears to exhibit fast local on most degenerate problems of the type we consider in this paper. Rather than our idea of a strict working set, SNOPT uses the slightly different concept of a *working set* $\mathcal{W}^k$ associated with each iteration k, with the properties that equality holds for the ith linearized constraint if $i \in \mathcal{W}^k$; the multiplier estimates λ_i^k are zero for $i \notin \mathcal{W}^k$; and the gradients $\nabla g_i(x^k)$, $i \in \mathcal{W}^k$ are linearly independent. We conjecture that the good behaviour of SNOPT is due to the fact that the working sets $\mathcal{W}^k$ eventually become subsets of a set $\tilde{\mathcal{B}}$ with the properties (i) and (ii) above. Features of SNOPT that promote this behavior, besides the maintenance of linear independence already mentioned, include

(a)   use of warm starts; that is, the working set $\mathcal{W}^k$ from the QP subproblem at iteration k is used as a starting estimate of the working set $\mathcal{W}^{k+1}$ at iteration $k + 1$; and

(b)   the fact that it allows constraints *not* in the working set ($i \notin \mathcal{W}^k$) to be violated by small tolerances.

Typical behavior of an algorithm with these properties is as follows. Because of linear independence of the gradients $\nabla g_i(x^k)$, $i \in \mathcal{W}^k$, a sufficiently advanced iterate will produce a working set $\mathcal{W}^k$ with the property (i). Iterate $k + 1$ then uses $\mathcal{W}^k$ as a starting guess and solves a QP that takes just the constraints $i \in \mathcal{W}^k$ into account. It finds a solution with this working set in which the "ignored" constraints $i \notin \mathcal{W}^k$ are violated only by small amounts, if at all, making this QP solution an acceptable approximation to the true SQP step. It then sets $\mathcal{W}^{k+1} = \mathcal{W}^k$, or possibly drops the

indices from $\mathcal{W}^{k+1}$ that have become inactive in the subproblem. The new working set usually retains the property (i), and subsequent iterations will stay with this set or some subset thereof, forcing the Lagrange multipliers to converge to a unique limit.

Unfortunately, a rigorous theoretical result concerning the SQP behavior just discussed does not seem to be possible. Instead, we propose here a formal algorithm called SQPsws (for "SQP with strict working sets") that is motivated by the informal procedure with warm starts and tolerances just described. We show that if, near the solution, a strict working set with the properties (i), (ii) is encountered at some iteration, then Algorithm SQPsws converges superlinearly thereafter, even if the strict working sets at later iterations fail to have one of these properties. While a strict working set of this type is likely to be identified in most practical situations, we can prove a rigorous convergence result only under similar conditions to those of the preceding two sections.

The key features of Algorithm SQPsws are its use of a stack of candidate warm-start strict working sets (instead of just the working set from the previous QP sub-problem) and its use of specific tolerances for the ignored constraints that are loose enough to allow the Lagrange multipliers to stabilize (and therefore to allow inexact solutions of the SQP subproblem) while being tight enough to guarantee superlinear convergence. Specifically, the tolerances require that violation of the ignored constraints be no more than $\eta(z^k, \lambda^k)^{1+\tau}$, where the quantity $\eta(\cdot, \cdot)$ defined in (6.5) measures distance from (z^k, λ^k) to the solution set $\mathcal{S}$, and τ is a parameter in the range $(0, 1)$. The stack of active subsets maintained by Algorithm SQPsws has the form

$$\texttt{top} \to \hat{\mathcal{B}}_s \to \hat{\mathcal{B}}_{s-1} \to \cdots \to \hat{\mathcal{B}}_1 \to \hat{\mathcal{B}}_0 = \{1, 2, \ldots, m\},$$

where s is a counter of stack size, the top element $\hat{\mathcal{B}}_s$ is the strict working set $\mathcal{B}^{k-1}$ from the previous iteration, and

$$\hat{\mathcal{B}}_s \subset \hat{\mathcal{B}}_{s-1} \subset \cdots \subset \hat{\mathcal{B}}_1 \subset \hat{\mathcal{B}}_0,$$

where all inclusions are strict. The index sets $\hat{\mathcal{B}}_{s-1}, \ldots, \hat{\mathcal{B}}_1$ are all strict working sets from previous iterations of the algorithm. Elements of the stack are popped and discarded if the solution to the subproblem (8.1) fails to meet the prescribed tolerances for the ignored constraints. As a last resort, if the stack is popped down to its last element $\hat{\mathcal{B}}^0$, the full SQP subproblem (4.1) is solved. In any case, the step produced by each iteration of the algorithm fits the iSQP framework (4.3), so the theory developed in Section 4 can be applied once again.

Algorithm SQPsws
choose $\tau \in (0, 1)$ and set (z^0, λ^0);
set $k \leftarrow 0$, $s \leftarrow 0$, $\hat{\mathcal{B}}_0 \leftarrow \{1, 2, \ldots, m\}$;
repeat
    set $\mu_k \leftarrow \eta(z^k, \lambda^k)$, `isqpsol` $\leftarrow$ `false`;
    while not `isqpsol`
        $\bar{\mathcal{B}} \leftarrow \hat{\mathcal{B}}_s$;
        solve the SQP subproblem for the constraint set $\bar{\mathcal{B}}$:

(8.1)
$$\min_{\Delta z} \Delta z^T \nabla \phi(z^k) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z^k, \lambda^k) \Delta z,$$

$$\text{subject to } g_i(z^k) + \nabla g_i(z^k)^T \Delta z \leq 0, \text{ all } i \in \bar{\mathcal{B}},$$

> denoting its strict working set by $\mathcal{B}^k$ and its primal-dual
> solution by $(\Delta z, \tilde{\lambda}_{\bar{\mathcal{B}}})$;
> **if** $g_i(z^k) + \nabla g_i(z^k)^T \Delta z \leq \mu_k^{1+\tau}$ for all $i \notin \bar{\mathcal{B}}$
>   isqpsol $\leftarrow$ true;
> **else**
> $s \leftarrow s - 1$;
> **end while**
> **if** $\mathcal{B}^k \neq \bar{\mathcal{B}}$
> $s \leftarrow s + 1, \hat{\mathcal{B}}_s \leftarrow \mathcal{B}^k$;
> $z^{k+1} \leftarrow z^k + \Delta z, \lambda^{k+1} \leftarrow (\tilde{\lambda}_{\bar{\mathcal{B}}}, 0)$;
> $k \leftarrow k + 1$;

until convergence.

Our analysis of Algorithm SQPsws requires a number of technical results, most of which pertain to the adequacy of the chosen subset $\bar{\mathcal{B}}$ of constraints in (8.1) in defining an approximate solution of the full subproblem (4.1), and to the progress that the resulting step makes toward the solution of the original problem (1.1).

We start by formalizing some of the ideas mentioned at the start of this section, which pertain to variants of the nonlinear programming problem (1.1) and the iSQP subproblem (4.3) in which some of the constraints are ignored. In (3.3), we defined the extended nonlinear programming problem $\text{NLP}(\tilde{\mathcal{B}})$ in which just a subset $\tilde{\mathcal{B}} \subset \mathcal{B}$ is enforced. We now define the extended iSQP subproblem corresponding to $\tilde{\mathcal{B}}$ (and to the extended nonlinear programming problem $\text{NLP}(\tilde{\mathcal{B}})$) as follows:

$$(8.2) \qquad \min_{\Delta z} \Delta z^T (\nabla \phi(z) + t) + \tfrac{1}{2} \Delta z^T \mathcal{L}_{zz}(z, \lambda) \Delta z,$$

$$\text{subject to } g_i(z) + \nabla g_i(z)^T \Delta z + r_i \leq 0, \ \ i \in \tilde{\mathcal{B}}.$$

The KKT conditions for (8.2) are

$$(8.3) \qquad \begin{bmatrix} \mathcal{L}_{zz}(z,\lambda)\Delta z + \nabla \phi(z) + t + \nabla g_{\tilde{\mathcal{B}}}(z)\lambda_{\tilde{\mathcal{B}}}^+ \\ g_{\tilde{\mathcal{B}}}(z) + \nabla g_{\tilde{\mathcal{B}}}(z)^T \Delta z + r_{\tilde{\mathcal{B}}} \end{bmatrix} \in \begin{bmatrix} 0 \\ N(\lambda_{\tilde{\mathcal{B}}}^+) \end{bmatrix}.$$

Note that (8.2) is truly an iSQP subproblem for (3.3) only if the Lagrangian $\mathcal{L}(z,\lambda)$ does not contain terms in its summation for indices i outside of the set $\tilde{\mathcal{B}}$, that is, only if $\lambda_i = 0$ for all $i \notin \tilde{\mathcal{B}}$. For generality, however, we allow $\lambda_{\{1,\ldots,m\}\setminus\tilde{\mathcal{B}}} \neq 0$ in some of the results below.

Our first technical result is a simple result based on Hoffman's lemma concerning the nearness of a given vector $\lambda \in \mathsf{R}^m$ (with property $\lambda_i = 0$, for all $i \notin \tilde{\mathcal{B}}$) to the unique optimal multiplier $\lambda^*(\tilde{\mathcal{B}})$ of $\text{NLP}(\tilde{\mathcal{B}})$ for some $\tilde{\mathcal{B}} \in \bar{\Phi}$.

LEMMA 8.1. *There is a constant $\beta \geq 1$ such that the following statement holds. For all $\lambda \in \mathsf{R}^m$ with the property that $\lambda_i = 0$ for all $i \notin \tilde{\mathcal{B}}$, for some $\tilde{\mathcal{B}} \in \bar{\Phi}$, we have*

$$\|\lambda - \lambda^*(\tilde{\mathcal{B}})\| = \|\lambda_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\| \leq \beta\delta(\lambda),$$

where, as always, $\delta(\lambda)$ denotes the distance from λ to the optimal multiplier set $\mathcal{S}_\lambda$ of the original problem (1.1).

Proof. Denoting by $P(\lambda)$ the closest vector in $\mathcal{S}_\lambda$ to λ, we have that

$$\delta(\lambda)^2 = \|\lambda - P(\lambda)\|^2 = \sum_{i\in\tilde{\mathcal{B}}}[\lambda_i - P(\lambda)_i]^2 + \sum_{i\in\mathcal{B}\setminus\tilde{\mathcal{B}}} P(\lambda)_i^2,$$

implying that

$$(8.4) \qquad \left\| P(\lambda)_{\mathcal{B} \backslash \tilde{\mathcal{B}}} \right\| \leq \delta(\lambda).$$

Note that $P(\lambda)$ satisfies the following system of linear equalities and inequalities:

$$(8.5) \qquad \sum_{i \in \mathcal{B}} \nabla g_i(z^*) P(\lambda)_i = \nabla \phi(z^*); \qquad P(\lambda) \geq 0; \qquad P(\lambda)_i = 0, \; i \notin \mathcal{B}.$$

The following system, on the other hand, has the unique solution $\bar{\lambda} = \lambda^*(\tilde{\mathcal{B}})$:

$$(8.6) \qquad \sum_{i \in \mathcal{B}} \nabla g_i(z^*) \bar{\lambda}_i = \nabla \phi(z^*); \qquad \bar{\lambda} \geq 0; \qquad \bar{\lambda}_i = 0, \; i \notin \tilde{\mathcal{B}}.$$

Because of (8.5), we have that $P(\lambda)$ violates (8.6) only in that possibly $P(\lambda)_i > 0$ for some $i \in \mathcal{B} \backslash \tilde{\mathcal{B}}$. Hence, by Hoffman's lemma [13] and uniqueness of $\lambda^*(\tilde{\mathcal{B}})$, we find that there is a number $\beta(\tilde{\mathcal{B}})$ such that

$$\| \lambda^*(\tilde{\mathcal{B}}) - P(\lambda) \| \leq \beta(\tilde{\mathcal{B}}) \left\| P(\lambda)_{\mathcal{B} \backslash \tilde{\mathcal{B}}} \right\|.$$

By choosing

$$\beta = \max_{\tilde{\mathcal{B}} \in \bar{\Phi}} \beta(\tilde{\mathcal{B}}) + 1,$$

combining this expression with (8.4), and using $\| P(\lambda) - \lambda \| = \delta(\lambda)$, we have that

$$\| \lambda - \lambda^*(\tilde{\mathcal{B}}) \| \leq \| \lambda - P(\lambda) \| + \| P(\lambda) - \lambda^*(\tilde{\mathcal{B}}) \| \leq \beta \delta(\lambda),$$

giving the result. $\blacksquare$

We now modify two of the results of Section 5 to apply to those extended problems (8.2) for which z^* satisfies Condition 2s.1, that is, $\tilde{\mathcal{B}} \in \bar{\Phi}$. The combination of these two results—Lemma 5.1 and Theorem 5.3—with Lemma 8.1 yields some powerful estimates.

LEMMA 8.2. *Suppose that Assumption 1, Condition 2s.1, and MFCQ hold. Then there exists a threshold value $\bar{\delta} > 0$ with the following property. If $\delta(z, \lambda) \leq \bar{\delta}$ and $\| (t, r_{\tilde{\mathcal{B}}}) \| \leq \bar{\delta}$ and $\lambda_i = 0$ for all $i \notin \tilde{\mathcal{B}}$ for some $\tilde{\mathcal{B}} \in \bar{\Phi}$, then the extended iSQP subproblem (8.2) has at least one solution $(\Delta z, \lambda^+)$, and for all such solutions we have*

$$(8.7) \qquad \| \Delta z \| + \| \lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}}) \| = O(\| z - z^* \|) + O(\| (t, r_{\tilde{\mathcal{B}}}) \|),$$

where $\lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})$ is the (unique) optimal Lagrange multiplier for NLP($\tilde{\mathcal{B}}$) (3.3). Moreover, we have that

$$\| z + \Delta z - z^* \| + \| \lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}}) \| \leq \| \lambda_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}}) \| O(\delta(z)) + O(\delta(z)^2) + O(\| (t, r_{\tilde{\mathcal{B}}}) \|)$$

$$= O(\delta(\lambda)\delta(z)) + O(\delta(z)^2) + O(\| (t, r_{\tilde{\mathcal{B}}}) \|)$$

$$(8.8) \qquad\qquad = O(\delta(z, \lambda)^2) + O(\| (t, r_{\tilde{\mathcal{B}}}) \|).$$

Proof. For a given $\tilde{\mathcal{B}} \in \bar{\Phi}$, we obtain by applying Lemma 5.1 to (8.3) that there is a threshold $\bar{\delta}(\tilde{\mathcal{B}})$ such that (8.7) holds whenever

$$(8.9) \qquad \left\| \begin{bmatrix} z - z^* \\ \lambda_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}}) \end{bmatrix} \right\| \leq \bar{\delta}(\tilde{\mathcal{B}}), \quad \|(t, r_{\tilde{\mathcal{B}}})\| \leq \bar{\delta}(\tilde{\mathcal{B}}).$$

(Note that we can write the distance of $\lambda_{\tilde{\mathcal{B}}}^+$ to the dual solution set for (3.3) explicitly as $\|\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\|$, since this set contains just the single element $\lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})$.) By Lemma 8.1, we have for all (z, λ) with $\lambda_i = 0$, $i \notin \tilde{\mathcal{B}}$ that

$$\left\| \begin{bmatrix} z - z^* \\ \lambda_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}}) \end{bmatrix} \right\| \leq \|z - z^*\| + \beta\delta(\lambda) \leq 2\beta\delta(z, \lambda),$$

where the last inequality follows from $\beta \geq 1$. It follows that both bounds in (8.9) are satisfied if we set

$$\bar{\delta} = \frac{1}{2\beta} \min_{\tilde{\mathcal{B}} \in \bar{\Phi}} \bar{\delta}(\tilde{\mathcal{B}}).$$

For (8.8), we obtain by applying Theorem 5.3 to the extended iSQP problem (8.2) that

$$\|z + \Delta z - z^*\| + \|\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\| \leq \|\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}\|O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|).$$

By applying the triangle inequality to the term $\|\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}\|$, reducing $\bar{\delta}$ if necessary to ensure that the $O(\delta(z))$ term is smaller than $1/2$, and rearranging, we obtain that

$$\|z + \Delta z - z^*\| + (1/2)\|\lambda_{\tilde{\mathcal{B}}}^+ - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\| \leq \|\lambda_{\tilde{\mathcal{B}}} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\|O(\delta(z)) + O(\delta(z)^2) + O(\|(t, r_{\tilde{\mathcal{B}}})\|),$$

yielding the first inequality in (8.8). The second relation follows immediately from Lemma 8.1, while the third follows from (2.12). $\square$

We now are in a position to prove our main convergence result for Algorithm SQPsws. We show that if a strict working set $\mathcal{B}^k$ from $\bar{\Phi}$ enters the stack at some sufficiently advanced iterate, then it remains in the stack and the algorithm converges superlinearly.

THEOREM 8.3. *Suppose that Assumption 1, Condition 2s.1, and MFCQ hold. Then there exists a positive threshold $\hat{\delta}$ such that if at some iteration $\bar{k}$ we have $\delta(z^{\bar{k}}, \lambda^{\bar{k}}) \leq \hat{\delta}$, and if there is an index set $\tilde{\mathcal{B}} \in \bar{\Phi}$ present in the stack at the start of iteration $\bar{k}$, then $\tilde{\mathcal{B}}$ remains in the stack at all subsequent iterations, and Algorithm SQPsws converges superlinearly with Q-order $1 + \tau$.*

Proof. All subproblems (8.1) considered by Algorithm SQPsws in which the conditions $g_i(z^k) + \nabla g_i(z^k)^T \Delta z \leq \mu_k^{1+\tau}$, all $i \notin \tilde{\mathcal{B}}$ have the form of (4.3) when we set $t = 0$ and r to be a vector with elements whose magnitude does not exceed $\eta(z^k, \lambda^k)^{1+\tau}$. Because of (6.5), we can choose $\hat{\delta}$ small enough that $\delta(z^k, \lambda^k) \leq \hat{\delta}$ implies that

$$\delta(z^k, \lambda^k) \leq \bar{\delta}, \quad \|(0, r)\| \leq \bar{\delta},$$

where $\bar{\delta}$ is the threshold value such that the assumptions of of Lemma 5.1, Theorem 5.2, and Theorem 5.3 are satisfied when $\delta(z, \lambda) \leq \bar{\delta}$ and $\|(t, r)\| \leq \bar{\delta}$. We reduce $\hat{\delta}$ if necessary so that the conditions of Lemma 8.2 are satisfied by (z, λ), $t = 0$, and

$r_{\mathcal{B}} = O(\eta(z,\lambda)^{1+\tau})$ whenever $\delta(z^k,\lambda^k) \leq \hat{\delta}$. Further assumptions on the size of $\hat{\delta}$ are made in the course of the proof.

The main part of our proof is to show that at any iterate k for which

(8.10a) $\tilde{\mathcal{B}}$ is present in the stack at the start of iteration k; and

(8.10b) $\delta(z^k,\lambda^k) \leq \hat{\delta},$

we have that

(8.11a) $\tilde{\mathcal{B}}$ is present in the stack at the end of iteration k;

(8.11b) $\lambda^{k+1}_{\{1,\dots,m\}\setminus\tilde{\mathcal{B}}} = 0$; and

(8.11c) $\delta(z^{k+1},\lambda^{k+1}) = O(\delta(z^k,\lambda^k)^{1+\tau}) \leq \delta(z^k,\lambda^k).$

By definition, the premise (8.10) is satisfied at iteration $\bar{k}$, and so from (8.11a) and (8.11c), it holds for all subsequent iterations. Hence, (8.11c) implies superlinear convergence.

Suppose that (8.10) holds for some k. Since $\tilde{\mathcal{B}}$ is present in the stack at the start of this iteration, the active set $\mathcal{B}^{k-1}$ from the subproblem (8.1) at the previous iteration must be such that $\mathcal{B}^{k-1} \subset \tilde{\mathcal{B}}$. In particular, we have that $\lambda^k_i = 0$ for all $i \notin \tilde{\mathcal{B}}$.

At the end of iteration k, the set $\tilde{\mathcal{B}}$ can have disappeared from the stack only if it was tried and rejected in (8.1), that is, if the solution to (8.1) obtained with $\bar{\mathcal{B}} = \tilde{\mathcal{B}}$ had

(8.12) $g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k > \mu_k^{1+\tau},$ for some $i \notin \tilde{\mathcal{B}}.$

Because of our choice of $\hat{\delta}$, we can apply Lemma 8.2 to (8.1) by setting $\bar{\mathcal{B}} = \tilde{\mathcal{B}}$, $(z,\lambda) = (z^k,\lambda^k)$, and $(t,r_{\tilde{\mathcal{B}}}) = 0$. We obtain from (8.8) that

$$\|z^k + \Delta z^k - z^*\| = O(\delta(z^k,\lambda^k)^2),$$

while from (8.7), we have

(8.13) $\|\Delta z^k\| = O(\delta(z^k)) = O(\delta(z^k,\lambda^k)).$

Hence, because of $g_{\mathcal{B}}(z^*) = 0$, we have that

$$
\begin{aligned}
g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k &= g_i(z^k + \Delta z^k) + O(\|\Delta z^k\|^2) \\
&= O(\|z^k + \Delta z^k - z^*\|) + O(\|\Delta z^k\|^2) \\
&= O(\delta(z^k,\lambda^k)^2), \qquad \text{for all } i \in \mathcal{B}.
\end{aligned}
$$

Since from (6.5) we have that $\mu_k = \Theta(\delta(z^k,\lambda^k))$, the condition (8.12) cannot hold for any $i \in \mathcal{B}$. Neither can it hold for any inactive constraint for the NLP, because by choosing $\hat{\delta}$ small enough, we have from (8.13) that

$$g_i(z^k) + \nabla g_i(z^k)^T \Delta z^k \leq (1/2)g_i(z^*) < 0, \qquad \text{for all } i \notin \mathcal{B}.$$

Hence, the violation (8.12) does not occur, so the set $\tilde{\mathcal{B}}$ will not be popped from the stack.

Since $\tilde{\mathcal{B}}$ remains in the stack, we must have $\mathcal{B}^k \subset \tilde{\mathcal{B}}$, so that (8.11b) holds.

Since $\mathcal{B}^k \subset \tilde{\mathcal{B}}$, we have that the primal-dual solution of (8.1) with $\bar{\mathcal{B}} = \mathcal{B}^k$ is an approximate solution to the extended iSQP subproblem (8.2) with

$$t = 0, \qquad r_{\mathcal{B}^k} = 0, \qquad 0 \leq -r_i \leq \mu_k^{1+\tau} \text{ for } i \in \tilde{\mathcal{B}}\setminus\mathcal{B}^k.$$

Because of our assumptions on $\hat{\delta}$, Lemma 8.2 applies to (8.1) with $\bar{\mathcal{B}} = \mathcal{B}^k$, and we obtain from (8.8) that

$$\|z^k + \Delta z^k - z^*\| + \|\lambda_{\tilde{\mathcal{B}}}^{k+1} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\| = O(\mu_k^{1+\tau}) = O(\delta(z^k, \lambda^k)^{1+\tau}).$$

Since

$$\delta(z^k + \Delta z^k, \lambda^{k+1}) \leq \|z^k + \Delta z^k - z^*\| + \|\lambda_{\tilde{\mathcal{B}}}^{k+1} - \lambda_{\tilde{\mathcal{B}}}^*(\tilde{\mathcal{B}})\|,$$

the first relation in (8.11c) follows. The second relation in (8.11c) follows from a sufficiently small choice of $\hat{\delta}$. $\square$

It seems reasonable to expect a set from $\bar{\Phi}$ to enter the stack at some sufficiently advanced iteration in most nonpathological cases. The strict working set $\mathcal{B}^k$ (for (z^k, λ^k) close to $\mathcal{S}$) is likely to belong at least to Φ by the following argument: The solution of (8.1) satisfies at least the second-order necessary conditions for the quadratic subproblem in which just the constraints $i \in \mathcal{B}^k$ are enforced. Since we know from Lemma 4.1 that there exists at least one $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subset \mathcal{B}^k$, we can reasonably expect that at least second-order necessary conditions are satisfied at z^* for $\text{NLP}(\mathcal{B}^k)$ as well. Hence, we would have $\mathcal{B}^k \notin \Phi$ only if the second-order conditions for $\text{NLP}(\mathcal{B}^k)$ are necessary but not sufficient—an uncommon scenario. Moreover, we can expect $\mathcal{B}^k$ to belong to the more restricted set $\bar{\Phi}$ because, as mentioned above, active-set solvers for the quadratic subproblem typically ensure that the active constraint Jacobian $\nabla g_{\mathcal{B}^k}(z^k)$ has full rank. (This property can be assured in any case by the simple procedure below.) In fact, our toleration of small violations in the non-enforced constraints $i \notin \mathcal{B}^k$ tends to discourage even nearly-dependent active constraint sets. Hence, the optimal Lagrange multiplier vector corresponding to $\mathcal{B}^k$ will be unique unless $\nabla g_{\mathcal{B}^k}$ loses rank between z^k and z^*—another uncommon occurrence.

The preceding paragraph suggests that we can prove that the conditions of Theorem 8.3 are satisfied if we make a few additional assumptions. One such assumption— full rank of $\nabla g_{\mathcal{B}^k}(z^k)$—can be guaranteed by applying a procedure based on the following observations to remove some indices from $\mathcal{B}^k$ if necessary. Any solution $(\Delta z^k, \lambda^{k+1})$ of (8.1) satisfies the system

$$(8.14) \qquad \sum_{i \in \mathcal{B}^k} \nabla g_i(z^k) \lambda_i^{k+1} = -\nabla \phi(z^k) - \mathcal{L}_{zz}(z^k, \lambda^k) \Delta z^k.$$

If $\nabla g_{\mathcal{B}^k}(z^k)$ does not have full rank, there is a vector $\Delta \lambda \neq 0$ such that

$$\sum_{i \in \mathcal{B}^k} \nabla g_i(z^k) \Delta \lambda_i = 0.$$

Because of the MFCQ condition (2.10), we have for z^k sufficiently close to z^* that at least one component of $\Delta \lambda$ is negative. Since by (8.14), the vector $(\Delta z^k, \lambda^{k+1} + \alpha \Delta \lambda)$ is a primal-dual solution of (8.1) for all α such that $\lambda^{k+1} + \alpha \Delta \lambda \geq 0$, we can choose α to reduce at least one component of $\lambda^{k+1} + \alpha \Delta \lambda$ to zero. By applying this procedure repeatedly as needed, we can arrive at a revised strict working set $\mathcal{B}^k$ with the desired property. In fact, by allowing a small violation of the equality in (8.14)—a violation $t = O(\mu_k^{1+\tau})$ that stays within the iSQP framework (4.4) and hence retains the stated convergence rate—we can remove even nearly-dependent constraints from the active set $\mathcal{B}^k$, thus increasing the likelihood that $\mathcal{B}^k$ belongs to $\bar{\Phi}$.

COROLLARY 8.4. *Suppose that Assumption 1, Condition 2s.3, and MFCQ are satisfied and that the CRCQ condition of Janin [14] holds; that is, there is an open*

neighborhood of z^* such that for any subset $\hat{\mathcal{B}}$ of $\mathcal{B}$, the matrix $[\nabla g_i(z)]_{i \in \hat{\mathcal{B}}}$ has constant rank for all z in this neighborhood. Then there exists a positive threshold $\bar{\delta}$ such that if $\delta(z^k, \lambda^k) \leq \bar{\delta}$ for some k, Algorithm SQPsws converges superlinearly.

Proof. We prove the result by showing that if $\delta(z^k, \lambda^k) \leq \bar{\delta}$ for some point (z^k, λ^k), then solution of an iSQP subproblem at this point will yield an active set $\mathcal{B}^k$ for which $\mathcal{B}^k \in \bar{\Phi}$. Since we know that *every* iteration of Algorithm SQPsws takes a step that fits the iSQP framework (4.3), it follows that $\mathcal{B}^k$ is pushed onto the stack at this point. Hence, superlinear convergence follows from Theorem 8.3.

From Lemma 4.1, we have that the strict working set $\mathcal{B}^k$ generated by iteration k is such that there exists at least one $\lambda^* \in \mathcal{S}_\lambda$ with $\mathcal{B}_+(\lambda^*) \subset \mathcal{B}^k$. Hence, by Lemma 3.3, we have that $\mathcal{B}^k \in \Phi$. Since by our discussion above, the active constraint Jacobian $\nabla g_{\mathcal{B}^k}(z^k)$ has full rank, and since the constant rank condition holds, we have that $\nabla g_{\mathcal{B}^k}(z^*)$ has full rank also. Therefore by Lemma 3.2, we have $\mathcal{B}^k \in \bar{\Phi}$. $\square$

Note that the constant-rank condition assumed here is stronger than the corresponding condition (7.8) used by Fischer [8].

Acknowledgments. I thank Michael Wagner, Mihai Anitescu, Alex Shapiro, Philip Gill, and Andreas Fischer for many interesting discussions on this topic. Thanks also to Michael for for carrying out numerical experiments during his summer at Argonne in 1997. I am also grateful to two referees for their comments on the first version and their pointers to related literature, which improved this version of the paper considerably.

Appendix A. Estimating the Distance to the Optimal Set.

An estimate of the distance from the current point (z, λ) to the primal-dual optimal set $\mathcal{S}$ is a critical ingredient in the modifications to the SQP algorithm discussed above. We show here that the easily computed quantity $\eta(z, \lambda)$ (6.5) is a satisfactory estimate in a neighborhood of $\mathcal{S}$.

THEOREM A.1. *Suppose that Assumption 1, Condition 2s.1, and the MFCQ condition is satisfied. Then if $\lambda \geq 0$, we have that*

$$\eta(z, \lambda) \stackrel{\text{def}}{=} \left\| \begin{bmatrix} \mathcal{L}_z(z, \lambda) \\ \min(\lambda, -g(z)) \end{bmatrix} \right\| = \Theta(\delta(z, \lambda)).$$

Proof. We start with the easy part of the proof, which is to show that $\eta(z, \lambda) = O(\delta(z, \lambda))$. By the assumed smoothness of ϕ and g, we have

$$(A.1) \qquad \mathcal{L}_z(z, \lambda) = \mathcal{L}_z(z, \lambda) - \mathcal{L}_z(z^*, P(\lambda)) = O(\delta(z, \lambda)).$$

For $\delta(z, \lambda)$ sufficiently small, we have $0 \leq \lambda_i < -g_i(z)$ for all $i \notin \mathcal{B}$, and therefore

$$(A.2) \qquad i \notin \mathcal{B} \ \Rightarrow \ 0 \leq \min(\lambda_i, -g_i(z)) = \lambda_i = |\lambda_i - P(\lambda)_i| \leq \delta(\lambda).$$

For the active indices $i \in \mathcal{B}$, we have

$$(A.3) \ \ i \in \mathcal{B} \ \Rightarrow \ \min(\lambda_i, -g_i(z)) \leq \max(0, -g_i(z)) \leq |g_i(z) - g_i(z^*)| = O(\delta(z)).$$

The result $\eta(z, \lambda) = O(\delta(z, \lambda))$ follows from the estimates (A.1), (A.2), and (A.3).

The more difficult part of the proof is to show that $\delta(z, \lambda) = O(\eta(z, \lambda))$. Our main theoretical tool is again Theorem 4.2 of Robinson [17].

We first define the vectors $v \in \mathsf{R}^m$ and $\omega \in \mathsf{R}^m$ as follows:

$$(A.4) \qquad v_i = \begin{cases} -g_i(z) & \text{if } -g_i(z) < \lambda_i \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, 2, \ldots, m,$$

(A.5) $$\omega_i = \begin{cases} \lambda_i & \text{if } -g_i(z) \geq \lambda_i \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, 2, \ldots, m.$$

For each $i = 1, 2, \ldots, m$, we have that either $g_i(z) + v_i = 0$ or $\lambda_i - \omega_i = 0$. We have also that $g(z) + v \leq 0$ and $\lambda - \omega \geq 0$, and therefore, by definition of $N(\cdot)$ in (2.5), we have that

(A.6) $$g(z) + v \in N(\lambda - \omega).$$

Note too that v and ω are complementary; that is,

(A.7) $$v \geq 0, \quad \omega \geq 0, \quad v^T \omega = 0.$$

In fact, we have that

$$v + \omega = \min(-g(z), \lambda),$$

and so from (A.7), (A.2), and (A.3) we obtain that

(A.8) $$\|v\|^2 + \|\omega\|^2 = \|v + \omega\|^2 = \|\min(-g(z), \lambda)\|^2 = O(\delta(z, \lambda)^2).$$

We therefore have the following estimates for v and w:

(A.9) $$\|v\| \leq O(\delta(z, \lambda)), \qquad \|\omega\| \leq O(\delta(z, \lambda)).$$

We now define perturbed variants of the objective function ϕ and constraint function g as follows:

$$\hat{\phi}(\hat{z}; \hat{u}, \hat{v}) \overset{\text{def}}{=} \phi(\hat{z}) - \hat{z}^T \hat{u}$$
$$\hat{g}(\hat{z}; \hat{u}, \hat{v}) \overset{\text{def}}{=} g(\hat{z}) + \hat{v},$$

where $(\hat{u}, \hat{v})$ is the perturbation vector. Note that $\hat{\phi}(\cdot; 0, 0) = \phi(\cdot)$ and $\hat{g}(\cdot; 0, 0) = g(\cdot)$. It is not difficult to show, with the help of (A.6), that $(\hat{z}, \hat{\lambda}) = (z, \lambda - \omega)$ is a primal-dual solution of following perturbed version of (1.1):

$$\min_{\hat{z}} \hat{\phi}(\hat{z}; \hat{u}, \hat{v}) \qquad \text{subject to } \hat{g}(\hat{z}; \hat{u}, \hat{v}) \leq 0,$$

where

$$\hat{u} = \mathcal{L}_z(z, \lambda - \omega), \qquad \hat{v} = v.$$

Both perturbations are small. For $\hat{u}$, we have from (A.1) and (A.9) that

$$\|\hat{u}\| = \|\mathcal{L}_z(z, \lambda - \omega)\|$$
(A.10) $$\leq \|\mathcal{L}_z(z, \lambda)\| + \left\| \sum_{i=1}^{m} \omega_i \nabla g_i(z) \right\| \leq O(\delta(z, \lambda)) + O(\|\omega\|) = O(\delta(z, \lambda)),$$

while for $\hat{v}$, we have immediately from (A.9) that $\|\hat{v}\| = \|v\| = O(\delta(z, \lambda))$. Hence, $(z, \lambda - \omega)$ is the solution of a slightly perturbed nonlinear program, where the size of the perturbation is uniformly small for (z, λ) near $\mathcal{S}$, so we can apply Theorem 4.2 of Robinson [17].

By the first inequality in the cited theorem, we have that

$$\delta(z, \lambda - \omega) = O\left(\text{dist}\left(0, \begin{bmatrix} \mathcal{L}_z(z, \lambda - \omega) \\ g(z) \end{bmatrix} - \begin{bmatrix} 0 \\ N(\lambda - \omega) \end{bmatrix}\right)\right)$$

(A.11)
$$= O\left(\|\mathcal{L}_z(z, \lambda - \omega)\| + \text{dist}(g(z), N(\lambda - \omega))\right).$$

For the first term, we have as in (A.10) that

$$\|\mathcal{L}_z(z, \lambda - \omega)\| = O(\|\mathcal{L}_z(z, \lambda)\|) + O(\|\omega\|).$$

For the second term, we have by application of the triangle inequality and (A.6) that

$$\text{dist}(g(z), N(\lambda - \omega)) \leq \|v\| + \text{dist}(g(z) + v, N(\lambda - \omega)) = \|v\|.$$

By substituting these estimates into (A.11), we obtain

(A.12)
$$\delta(z, \lambda - \omega) = O\left(\|\mathcal{L}_z(z, \lambda)\| + \|v\| + \|\omega\|\right).$$

By applying the triangle inequality again, we obtain from (A.8) and (A.12) that

$$\begin{aligned}
\delta(z, \lambda) &\leq \delta(z, \lambda - \omega) + \|\omega\| \\
&= O\left(\|\mathcal{L}_z(z, \lambda)\| + \|v\| + \|\omega\|\right) \\
&= O\left(\|\mathcal{L}_z(z, \lambda)\| + \|\min(-g(z), \lambda)\|\right) \\
&= O(\eta(z, \lambda)),
\end{aligned}$$

as required. $\square$

The estimate (6.5) was proposed by several other authors independently of this paper. Facchinei, Fischer, and Kanzow propose the same estimate in a revised version of their paper [7]. Hager and Gowda [12, Theorem 1, Theorem 3] propose a more general measure, which reduces to (6.5) when $\lambda \geq 0$ and does not require the MFCQ condition to hold.

Appendix B. Perturbation Analysis of a Convex Program.
We consider the following convex quadratic program:

(B.1)
$$\min_x \tfrac{1}{2} x^T Q x + c^T x \quad \text{subject to} \ Ax = b, \ Cx \leq d,$$

where Q is symmetric positive definite. Suppose the constraints satisfy the following property

(B.2)
$$Aw = 0, \quad Cw < 0, \quad \text{for some vector } w.$$

If in addition we were to assume that the rows of A were linearly independent, these constraints would satisfy the MFCQ. We have the following result.

LEMMA B.1. *Consider the problem (B.1), where we take Q, A, and C to be fixed while c, b, and d are allowed to vary. Assume that (B.2) holds. Then*
   (i) *(B.1) has at most one solution $x(c, b, d)$ for any vector triple (c, b, d), and if in addition the rows of A are linearly independent it has exactly one solution;*
   (ii) *If the solution exists for two vector triples (c, b, d) and (c', b', d'), the following Lipschitz continuity property is satisfied:*

   (B.3)
   $$\|x(c, b, d) - x(c', b', d')\| \leq L\|(c, b, d) - (c', b', d')\|,$$

   where the constant L depends only on Q, A, and C.

Proof. Because the objective function is strictly convex and the feasible region is convex polyhedral, a unique solution will exist whenever the feasible region is nonempty. When the MFCQ is satisfied, the feasible set is in fact nonempty for all b and d. Therefore, (i) is true.

The proof of (ii) is similar to that of Proposition 7.5.9 and Corollary 7.5.10 of Cottle, Pang, and Stone [6], so we omit the details. ◻

REFERENCES

[1] M. C. BARTHOLOMEW-BIGGS, *Recursive quadratic programming methods for nonlinear constraints*, in Nonlinear Optimization 1981, M. J. D. Powell, ed., Academic Press, 1982, pp. 213–221.

[2] ———, *A globally convergent version of REQP for constrained minimization*, Technical Report 168, Hatfield Polytechnic, 1986.

[3] ———, *Recursive quadratic programming methods based on the augmented Lagrangian*, Mathematical Programming Study, 31 (1987), pp. 21–41.

[4] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[5] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, Mass., 2nd. ed., 1999.

[6] R. W. COTTLE, J.-S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, San Diego, 1992.

[7] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *On the accurate identification of active constraints*, SIAM Journal on Optimization, 9 (1998), pp. 14–32.

[8] A. FISCHER, *Modified Wilson method for nonlinear programs with nonunique multipliers*, Mathematics of Operations Research, 24 (1999), pp. 699–727.

[9] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Mathematical Programming, 12 (1977), pp. 136–138.

[10] P. E. GILL, W. MURRAY, AND M. SAUNDERS, *SNOPT: An SQP algorithms for large-scale constrained optimization*, Report NA 97-2, Department of Mathematics, University of California, San Diego, 1997.

[11] W. W. HAGER, *Stabilized sequential quadratic programming*, Computational Optimization and Applications, 12 (1999), pp. 253–273.

[12] W. W. HAGER AND M. S. GOWDA, *Stability in the presence of degeneracy and error estimation*, Mathematical Programming, Series A, 85 (1999), pp. 181–192.

[13] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, Journal of Research of the National Bureau of Standards, 49 (1952), pp. 263–265.

[14] R. JANIN, *Directional derivative of the marginal function in nonlinear programming*, Mathematical Programming Study, (1984), pp. 110–126.

[15] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz-John necessary optimality conditions in the presence of equality and inequality constraints*, Journal of Mathematical Analysis and Applications, 17 (1967), pp. 37–47.

[16] D. RALPH AND S. J. WRIGHT, *Superlinear convergence of an interior-point method despite dependent constraints*, Mathematics of Operations Research, 25 (2000), pp. 179–194.

[17] S. M. ROBINSON, *Generalized equations and their solutions. part II: Applications to nonlinear programming*, Mathematical Programming Study, 19 (1982), pp. 200–221.

[18] S. J. WRIGHT, *Superlinear convergence of a stabilized SQP method to a degenerate solution*, Computational Optimization and Applications, 11 (1998), pp. 253–275.

[19] S. J. WRIGHT AND D. RALPH, *A superlinear infeasible-interior-point algorithm for monotone nonlinear complementarity problems*, Mathematics of Operations Research, 21 (1996), pp. 815–838.